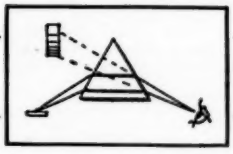
Journal of the OPTICAL SOCIETY *of* AMERICA

VOLUME 33
No. 9



SEPTEMBER
1943

Factors in Human Visual Resolution . . . . . GORDON L. WALLS 487

Some Aspects of the Eye as an Image-Forming Mechanism . . . . . KENNETH N. OGLE 506

Single Judgment Test for Red-Green Discrimination . . . . . LEGRAND H. HARDY 512

Improved Electrode Holder for Spectrographic Analysis . . . . . BOURDON F. SCRIBNER AND CHARLES H. CORLISS 515

Computation of Polynomial Functions by Summation of Finite Differences . H. F. BENNETT 519

Simplified Formulae for Scattered and Rescattered Sunlight . . . . . LUDWIK SILBERSTEIN 526

Letter to the Editor . . . . . E. O. HULBERT 505

Published monthly
at Prince and Lemon Sts., Lancaster, Pa.
by the
AMERICAN INSTITUTE OF PHYSICS
Incorporated
175 Fifth Avenue, New York, N. Y.
FOR AND UNDER THE EDITORIAL SUPERVISION OF THE
OPTICAL SOCIETY OF AMERICA

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

THE OPTICAL SOCIETY OF AMERICA

Officers of the Society

A. G. WORTHING, *President*
University of Pittsburgh
Pittsburgh, Pennsylvania

A. H. PFUND, *Vice President*
Johns Hopkins University
Baltimore, Maryland

ARTHUR C. HARDY, *Secretary*
Massachusetts Institute of Technology
Cambridge, Massachusetts

W. W. GRAEPER, *Treasurer*
Bausch & Lomb Optical Company
Rochester, New York

The Board of Directors consists of the above officers; the Editor of the Journal; the Past President; and five directors-at-large.

K. S. GIBSON, *Past President*
National Bureau of Standards
Washington, D. C.

GEORGE R. HARRISON, *Editor*
Massachusetts Institute of Technology
Cambridge, Massachusetts

Directors-at-Large

J. W. BEAMS
University of Virginia
Charlottesville, Virginia

D. B. JUDD
National Bureau of Standards
Washington, D. C.

RUDOLF KINGSLAKE
Eastman Kodak Company
Rochester, New York

D. L. MACADAM
Eastman Kodak Company
Rochester, New York

R. A. SAWYER
University of Michigan
Ann Arbor, Michigan

Rochester, New York, Section

J. H. WEBB, *President*
G. H. AKLIN, *Vice President*
A. F. TURNER, *Secretary*
F. H. PERRIN, *Treasurer*

Councillors:

F. E. ALTMAN
J. W. FORREST
F. M. E. HOLMES
V. L. PARKS
R. B. HORSFALL (*ex officio*)

Patrons

AMERICAN OPTICAL COMPANY — BAUSCH & LOMB OPTICAL COMPANY
EASTMAN KODAK COMPANY

Corporation Members

AMERICAN CYSTOSCOPE MAKERS, INC., New York, New York
AMERICAN OPTICAL COMPANY, Southbridge, Massachusetts
BAUSCH & LOMB OPTICAL COMPANY, Rochester, New York
BELL & HOWELL COMPANY, Chicago, Illinois
BELL TELEPHONE LABORATORIES, INC., New York, New York
BENDIX AVIATION CORPORATION, New York, New York
CENTURY LIGHTING, INC., New York, New York
CORNING GLASS WORKS, Corning, New York
E. I. DUPONT DE NEMOURS & COMPANY, INC., Chemical Department, Experimental Station, Wilmington, Delaware
EASTMAN KODAK COMPANY, Rochester, New York
FAIRCHILD AVIATION CORPORATION, Jamaica, New York
FISH-SCHURMAN CORPORATION, New York, New York
GAERTNER SCIENTIFIC CORPORATION, Chicago, Illinois
W. & L. E. GURLEY, Troy, New York
INTERCHEMICAL CORPORATION, New York, New York
KEUFFEL & ESSER COMPANY, Hoboken, New Jersey
NATIONAL CARBON COMPANY, Cleveland, Ohio
PERKIN-ELMER CORPORATION, Glenbrook, Connecticut
POLAROID CORPORATION, Cambridge, Massachusetts
PORTMAN MACHINE TOOL COMPANY, New Rochelle, New York
SIMPSON OPTICAL MANUFACTURING COMPANY, Chicago, Illinois
SPENCER LENS COMPANY, Buffalo, New York
ARTHUR H. THOMAS COMPANY, Philadelphia, Pennsylvania
ZENITH OPTICAL COMPANY, Huntington, West Virginia

THE OPTICAL SOCIETY OF AMERICA was organized in 1916 "*to increase and diffuse the knowledge of optics in all its branches, pure and applied, to promote the mutual interests of investigators of optical problems, of designers, manufacturers and users of optical instruments and apparatus of all kinds and to encourage cooperation among them.*" It aims to serve the interests of all who are engaged in any branch of optics from fundamental research to the manufacture of optical goods.

MEMBERSHIP

Regular Members: "Any person who has contributed materially to the advancement of optics is eligible for regular membership."

Associate Members: "Any person who is interested in optics is eligible for associate membership."

Corporate Members: "Any corporation desirous of encouraging the work of the Society is eligible for corporation membership."

Patrons: The payment, within any one year, of $1000 or more toward the expenses of the Society by any corporation or individual entitles such corporation or individual to the rank of patron.

The Society cordially invites to membership all who are interested in any branch of optics, either in research, in instruction, in optical or illuminating engineering, in the manufacture and distribution of optical goods of all kinds or in physiological and medical optics.

ANNUAL DUES

Regular members, $8.50; Associate members, $6.00; Corporate members, $50.00; Patrons, $1000.00. These dues include subscription to the Society's official journal, the *Journal of the Optical Society of America.*

CHANGES OF ADDRESS

All changes of address of members of the Optical Society of America should be communicated to the Secretary, Professor Arthur C. Hardy, Massachusetts Institute of Technology, Cambridge, Massachusetts.

Journal of the
OPTICAL SOCIETY *of* AMERICA

VOLUME 33, NUMBER 9 SEPTEMBER, 1943

Factors in Human Visual Resolution*

GORDON L. WALLS
Bausch and Lomb Optical Company, Rochester, New York

NO very precise definition of "visual resolution" could be given, but by this expression I mean the perception of pattern—in the broadest sense—in visual space. This performance is exceedingly complex, but there is a degree to which it is fairly safe to atomize it into participating capacities. Some of these fall under the heading of "visual acuity" as it is usually defined, while others do not. For, that term has unfortunately been preempted to describe a particular aspect of the whole performance of pattern-perception, namely the discrimination of two lines (or points) as two. Some would even insist that the lines must be *dark* ones. It is just as important, however, to see a line at all—bright or dark—and to see it where it is, as it is to see whether it is a single line or a double one.

There is a certain amount of confusion as regards just which of the various kinds of "acuity" one is calling upon in certain situations. This is particularly dangerous during this period of intensive experimentation with visual optical instruments, and can lead to false hopes and expectations regarding the performance of new devices and new modifications of old ones. It therefore seems desirable to review the processes which contribute to pattern-perception, and to check off the factors known to influence each of them.

* Modified from a paper presented at the Symposium on Vision at the meeting of the Optical Society of America in New York City, March 5–6, 1943.

VISIBILITY OF LINES AND POINTS

Bright-on-Dark

The visibility of a bright point or line against a dark background depends only upon its brightness. There is no limit, not even a theoretical one, to the smallness of an effective stimulus provided it is bright enough. Its image on the retina is certain to illuminate at least one visual cell, or one ragged line of visual cells, sufficiently to stimulate them. Only by coincidence will the true size of a just-visible bright object be perceived, however. Its apparent size will be determined by the size of that portion of its diffusion pattern which is bright enough to affect the visual cells in their current condition of adaptation. Where only one visual cell is stimulated, the object can appear constant in size over a considerable range of actual sizes and inversely corresponding brightnesses. Where the object is very small in actual size, brightening it will make it appear to grow, as the supraliminal portion of its retinal image reaches out over more and more rings of visual cells. This is the simple basis of the scale of star "magnitudes," which are brightnesses falsely perceived as sizes. A very bright star may appear to subtend several minutes of arc at the eye, since the effective size of its retinal image embraces several cone diameters; but the (angularly) largest star (Mira) subtends only 0.056″ at the earth.

BOSTON UNIVERSITY
COLLEGE OF LIBERAL ARTS
LIBRARY

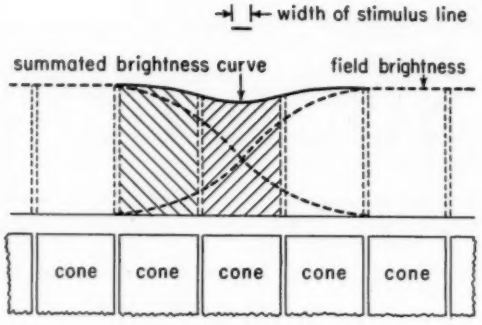


FIG. 1. Basis of dark-line visibility (modified from Moon). If the hatched areas (each representing the total illumination of the cone beneath) differ by one percent or more, the line will be seen.

Dark-on-Bright

The perception of a dark point or line upon a bright ground is more complex. This performance has been described as the perception of a difference in brightness between line and ground. More accurately, it is the perception of a difference between the background brightness and the sum of the fractions of that brightness which are displaced into the geometrical territory of the image, from either side of it, by diffraction at the pupil and the dioptric errors of the eye.

The dark line is not sharply imaged, but yields on the retina a shadow-like stripe in which the distribution of shadow-density is regulated primarily by diffraction and chromatic aberration—to a very minor extent also by spherical aberration. The center of the shadow is much darker than its borders; and if the line of retinal cones on which it falls receives a sufficient reduction of illumination (as compared with cones alongside, receiving the background brightness practically unchanged), the stimulus-line will be perceived—though not as a jet-black line if it is of not much more than threshold width.

If the retinal areas illuminated are not too small, the difference in their brightness which can be perceived falls rapidly from a very high value (100 percent or more) in scotopic vision until the cone threshold is reached, and thereafter more slowly, becoming practically constant at one percent over a considerable range of high intensities, and falling as low as 0.6 percent at very high intensities (Hecht, 1935). With small areas, no such low values of the relational threshold of brightness can be attained; but Hecht and Mintz (1939) take the attitude that in the perception of a long line, the total number of cones affected would, if aggregated compactly, cover an adequately-large area.

The basis of the perception of a dark line is diagrammed in Fig. 1. The oppositely-hatched areas represent the respective illuminations of adjacent cones, and must differ by (say) one percent if the stimulus-line is to be wide enough to see. We can ignore the negligible dimming of the adjacent rows of cones,[1] and imagine the penumbrous shadow on the central row of cones to be condensed into an absolute umbra. We should then have the situation in Fig. 2, wherein the completely dark stripe occupies one percent of the cross-sectional area of a cone, and has a width equal to 0.007854 times the cone diameter.

To all intents and purposes, this stripe is the geometrical optical image of the stimulus line. If we assume, for the subtense of a foveal cone, the reasonable value of 45″ of visual angle, the angular width of the stripe in Fig. 2 becomes 0.35″. Experimental values very close to this have indeed been obtained:

Barnard (1897–8), whose predecessors had all secured values of many seconds, placed a 0.0009-in. wire against the sky as a background and found that it could be "distinctly seen" at 356 feet, when its subtense was 0.44″ arc. Other astronomers interested in this same problem did not do quite as well. Lowell (1903) found that

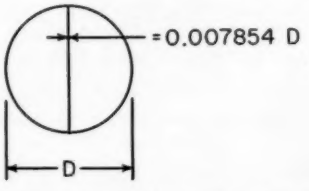


FIG. 2. Cross section of a foveal cone, with its moiety of the image of a just-visible line compressed into an absolute umbra, occupying one percent of the cross-sectional area. Further explanation in text.

[1] Lapicque's (1936) work indicates that for near-threshold observations with a pupil as large as 4 mm, the supraliminal portion of the retinal image of a bright point or line will be less than 4μ wide. It follows that the physiologically effective umbra of a dark point or line will be no wider.

an iron wire against the sky was continuously visible with ease at 1.13″, with difficulty at 0.83″, and visible intermittently at 0.73″. Pickering (1915), using a human hair, could see it easily at 1.15″ and intermittently at 0.82″. These findings might seem to throw suspicion upon Barnard's; but they probably only reflect the variability and unreliability of the sky background, discovered by Hecht and Mintz (1939) early in their work. These workers adopted, instead, wires stretched across a trans-illuminated opal glass disk two feet across and two or three meters away. Hecht could see a wire whose subtense was 0.5″, but did not regard this as the limit. The highest brightness used was only about 30 ml, which is near the bottom of the brightness range for the "constant," near-minimal relational threshold of one percent (see above). This brightness is, however, well within the range for the 4-mm "physiological" pupil which affords maximal visual acuity, other things being equal (see below).

It is interesting to compare some values obtained with dots. Pickering (1915) employed a pattern of seven India-ink dots, graded in size from 0.90 to 0.15 mm, on tracing cloth trans-illuminated by skylight. The respective sizes could be seen continuously at corresponding distances such that the subtenses grouped closely about an average of 33.6″. The average for "glimpsing" (i.e., intermittent visibility) was 30.1″. These values are many times the threshold widths for lines; but this is to be expected since a line can affect a great number of visual cells, in proportion to its length. A line of the width of Hecht and Mintz' least visible (0.5″) would need to be 30′ long, for its area to equal that of a circle 33.6″ in diameter.[2]

Hecht and Mintz' wires were never less than 11° 36′ in angular length. Whether they might still have been seen at a length of anything like 30′ is an interesting question. The brightness conditions of Pickering's and of Hecht and Mintz' situations were of course very different, and probably favored Pickering's observations decidedly. Lasareff (1911), working at an unspecified intensity, found that the relational threshold of brightness reached its "constant" and "minimal" value only with test fields as large as, or larger than, 40′ subtense.[3] If it is the *diameter* of this minimal test field which should be compared with the length of that single line of cones whose brightness-discrimination can be as low as one percent at high intensities, then Hecht and Mintz' wires were amply long—but at 30′ might have been too short. Suppose, however, that we take the 40′

FIG. 3. Diagram illustrating Wilcox and Purdy's viewpoint regarding the intergradation of two-line resolution and single-line visibility.

circular minimal area very literally: The greatest number of 45″ cones which could be packed into it would be 2234. Were these to be placed in single file to make a line equal to Lasareff's minimal area, the line would have an angular length of 28°—more than twice that of Hecht and Mintz' wires. It would seem that the field *diameter* fixes the range of the relational brightness threshold.

VISUAL ACUITY

Primary Considerations

The resolution of a pair of points—or, more usually, parallel lines—is the classical problem of "visual acuity." The minimal detail of the test pattern is converted to minutes of arc, and the reciprocal of this value then expresses the visual acuity. In the examination of such systems as photographic objectives, with such patterns as the Foucault test object, consisting of alternate black and white stripes of equal width, it is customary to take as the "minimal detail" the distance from the center of one black

[2] Duke-Elder (1933) gives the interesting observation that a row of dots can be seen when, individually, they are below visible size.

[3] More recently, Cobb and Moss (1928) and Heinz and Lippay (1928) have studied the effect of field size; but unfortunately the intensities used were not high enough for the relational threshold to reach minimum or "constancy" with any size of field used. The trend of Lasareff's findings was, however, confirmed.

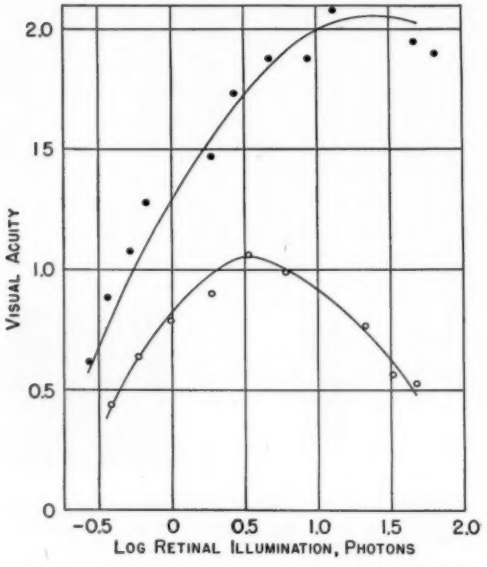


FIG. 4. Resolution of bright bars on a dark ground (open circles) and of dark bars on a bright ground (solid dots). Data from Wilcox and Purdy.

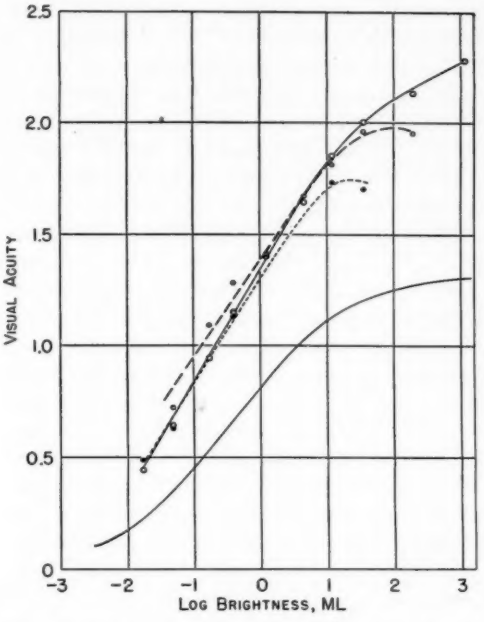


FIG. 5. Visual acuity in relation to intensity. Data for upper curves from Lythgoe's Table VI (subject "J. M."); Large circles (solid line), with surround brightness equal to target brightness, except at two highest points at which surround brightness was 36.3 ml; small circles (dashed line), with surround brightness 0.01 ml; solid dots (dotted line), with dark surround. Target for all three curves was a Landolt broken circle. Lower (pointless) curve represents the photopic portion of König's data, obtained with a Snellen hook target in a dark surround.

stripe to the center of the next. It is important to note that this was once customary in physiological optics, but is no longer—currently, the distance considered to be resolved is the width of the space between the lines. In reading accounts of performance in the older literature, one must watch out for instances in which the experimenter has expressed his thresholds in the photographer's terms, making the performance look only half as good as it really was. An occasional modern worker (e.g., Fabry, 1939) may also use the older expression.

Though the task in visual resolution is to perceive the gap, one's performance is influenced by the widths of the confining lines. The threshold width of the gap decreases when the lines are widened into bars (Roelofs, 1918; McFadden, 1940). Wilcox and Purdy (1933) point out that as the bars are widened indefinitely, one passes ultimately (no one seems to want to say when) from a resolution situation to that of a single line on a ground of anithetical brightness; and one has then a simple *visibility* task—of seeing a bright line on a dark, or a dark line on a bright, background (Fig. 3). The kinds of performance to be expected then have already been discussed above.

For two black lines of just-perceptible width, a theoretical limit of resolution can be calculated from the optical constants of the eye. Depending upon the choice of constants and upon the criterion of resolution (which is always a bit arbitrary), this limit will be from 36″ to one minute or more. Helmholtz (1867) chose 64″, which he considered to be the subtense of a foveal cone.

Until a few years ago, it was generally taught that resolution could be expected only if an unstimulated row of cones intervened between stimulated rows, for too many failed to notice that Helmholtz was talking about infinitely narrow lines. We have long known that foveal cones subtend something more like 45″ than 64″; but what makes us very sure that their dimen-

sions are not the limiting factor in resolution is the fact that thresholds as low as 25″ have been obtained experimentally with such test objects as the Snellen hook, the Landolt broken circle, or a single pair of parallel lines.[4]

The primary basis of resolution, as Helmholtz understood perfectly well, is always the distribution of brightness in the retinal image taken together with the relational threshold of brightness. This is readily grasped in the case of narrow bright lines on a dark field. Their diffusion-stripes on the retina must obviously be separated until at least one row of cones receives perceptibly less light than the rows on either side. If the brightness of the lines is low their separation must be increased, since at low illuminations the relational threshold of brightness is high. If, however, the brightness of the lines is made very high the separation must again be increased over the minimal, else the gap will be filled in by "irradiation": The retinal brightness in the gap rises more rapidly than that opposite the lines, quickly obliterating any perceptible difference between lines and gap. This could not occur if diffraction alone produced the diffusion of the geometrical-optical retinal image (for then, raising the brightness level would not alter the relative brightness in the image; see Hofmann, 1919), but chromatic aberration is also an important element in the diffusion; and there are factors other than strictly optical ones (Fry and Cobb, 1937). With wide bright bars, the fall in resolution at high intensities does not occur, for irradiation is lessened by any change in the field which reduces "scotomatic glare"—the *concentration* of the lighted areas. Central-nervous processes are held responsible for this difference (see Bartley, 1941).

The best performances reported with double stars, pairs of bright bars, and bright grilles are about one minute of arc (see Hartridge, 1922, 1923); but such results were never obtained until it had been realized that the subjective widening of the pattern units could be minimized only at an intermediate brightness. For, with dark lines on a bright field, quite another result is obtained: The brighter the field and surround, the smaller the perceptible separation of the lines (Fig. 4).[5]

For parallel black lines, photopic visual acuity was early concluded to increase in proportion to the logarithm of the background brightness over a wide range of intensities, as expressed in Fechner's "law." On the basis of König's (1897) classical work (Fig. 5), acuity was believed to fall off inevitably at high intensities. These data were interpreted by Hecht (1926, 1927, 1928) as indicating that the improvement of visual acuity with increased illumination had its basis in a successive involvement of entities whose values had a random distribution, plotting a "probability" curve. Hecht assumed that these entities were the thresholds of stimulation of the individual visual cells (though in the case of the rods, at least, a most conspicuous variable is the number of receptors from which each bipolar receives stimulation—influencing *its* commencement-of-activity). It has seemed unlikely to practically all other investigators that the cone thresholds are scattered over the entire range of photopic intensities which can be discriminated from neighboring intensities. Hecht (1930; 1934, p. 785) tries to minimize the difficulty by claiming that most of the variation occurs within less than two log units. Merely by

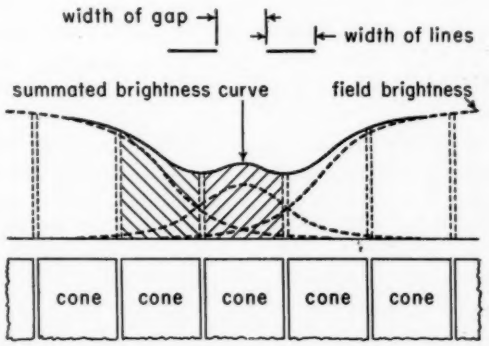


FIG. 6. Hartridge's conception of the basis of two-line resolution (modified from Moon). If the two oppositely-hatched areas (cf. Fig. 1) differ by one percent or more, the gap between the stimulus lines will be perceptible.

[4] With a grille ("grating"), such low thresholds can never be attained as with a single pair of lines. Helmholtz, and many an investigator since, employed grilles or—worse—grids.

[5] This should, theoretically, not be true for lines whose widths are adjusted, for each of the experimental intensities, so as to be just visible at that intensity. When lines wider than this are used, one is not—in a sense—dealing with threshold conditions at all when one makes the gap just discernible, and studies the effect of this or that upon the performance.

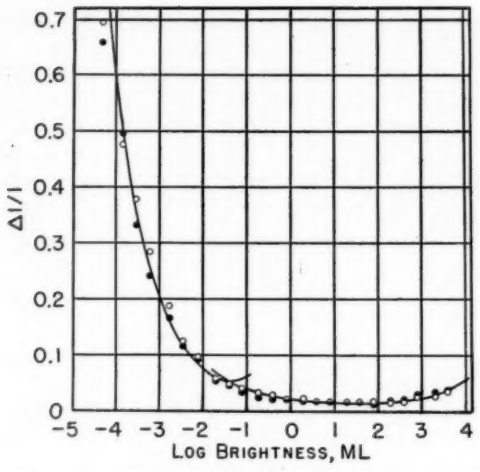


FIG. 7. Brightness discrimination in relation to intensity. Data from König and Brodhun (solid dots, Brodhun's eye; open circles, König's). The curves drawn through the data are those derived by Hecht from theoretical equations. Their crossing-point is not far away from the cone threshold.

choosing some unit of brightness larger than the millilambert, he could as easily have made it one log unit, or a tenth of a log unit—the expression of data in log units so notoriously obscures facts.

From the outset, Hecht placed the "recognition of intensity" and the discrimination of brightness upon this same basis of an increasing number of functional receptors with increasing intensity. The least noticeable difference in brightness, at any brightness level—falsely represented as "$\Delta I/I$" (see Moon, 1936)—is always determined by comparing *two* patches, one of which is brightened until it looks brighter than the other. Using this method, König and Brodhun (1888–9) estimated that there are 660 "steps" in the whole range of perceptibly-increasing brightness (see Wilcox and Purdy, 1933, p. 252). There are, however, no critical points—one can start one step anywhere, which in itself is a little embarrassing to Hecht's theory. Hecht laid great stress upon the discontinuity of brightness discrimination, and upon the particular discontinuity shown by König and Brodhun's particular method—perhaps forgetting that no one ever studies brightness discrimination by observing how much a single luminous patch has to be brightened before it can be seen to have brightened. One cannot "recognize" intensities. In vision, as in audition, there can be a "sense of absolute pitch" only for pitch—i.e., frequency.

The fact that "$\Delta I/I$" can be influenced by changes in field size (Lasareff, 1911; Cobb and Moss, 1928; Heinz and Lippay, 1928), when that size is already many times that of Hecht's "minimal retinal area" (containing only the few hundred cones necessary to provide a full assortment of cone thresholds), casts additional doubt upon Hecht's concept.

Hecht gave some consideration to differential frequency of discharge of the receptors themselves as a possible minor factor in the relation of brightness discrimination to brightness level; but his critics have pointed always to the well-known variability of *optic nerve-fiber* discharge frequency as the immediate basis (above the photochemical level of determination, that is) of brightness discrimination. It no longer seems necessary to suppose that the cone *thresholds* are widely spread, or that the cones—which are not nerve cells at all—obey the all-or-none law in any way. On the photochemical basis supplied by Hecht himself (and cf. Pütter 1918), each cone is capable of an almost infinitely graded series of photochemical stationary states or activity levels, which can express themselves in a series of discharge frequencies in the associated neurons.

In more recent discussions Hecht (e.g., 1934) has given room to these ideas, though he still clings to a discontinuously graduated response of the visual cells, and says: "The amount of photochemical change required for the cell to discharge an impulse varies with the threshold of the cell." This amount *is* the threshold. The number of functional receptors can become constant shortly above the threshold of the most sensitive, the activity of each receptor can be both continuous and smoothly graded, and brightness discrimination will still be discontinuous if nerve-impulse frequencies are discontinuously graded or discontinuously registered in consciousness. The rise of visual acuity with illumination may well be discontinuous also, though no one has reported that it is.

Hartridge (1922), foreshadowed by Hofmann (1919), brought Helmholtz' ideas up to date.

Brightness discrimination and distribution of brightness in the image are the sole factors in his conception, which is diagrammed in Fig. 6. The illumination of the central line of cones only needs to be perceptibly different from that of the adjacent lines, and the gap between the stimulus-lines will be perceived. That gap may of course subtend much less than one cone diameter. Since the apparent all-importance of brightness discrimination is so generally agreed upon, it behooves us to make sure that the effects of increasing intensity upon visual acuity and upon brightness discrimination do indeed run parallel:

Eguchi (1932) appears to have employed the highest intensities, and to have obtained the highest visual acuities. His test object was a Landolt broken circle on a transilluminated disk 20.5 cm in diameter at close range (in a room 4.1×3.2 m). He summarized his results as follows:

Ground brightness, ml	Threshold, seconds
1.22	35
15.9	24
1700	24
77500	24

Thus, visual acuity did not fall off as in König's experiments, which were done with a small target and a dark surround, and yielded a minimal threshold of 47″. Lythgoe (1932), partly anticipated by Cobb and Geissler (1913), demonstrated that when the visual field as a whole is kept almost or quite as bright as the target-background proper, visual acuity does not even cease to rise with intensity but continues to improve indefinitely as long as the subject can stand the heat of the field lamps alongside his head. Lythgoe's subject ("J. M.") reached a threshold which was slightly inferior to Eguchi's (i.e., 26″) with the same test-object; but then, Lythgoe did not use quite such high brightnesses (Fig. 5).

Such findings were disturbing to the strict constructionists, who had remained happy so long as both visual acuity and brightness discrimination were thought to plateau, and then worsen (Fig. 7—cf. König curve and Lythgoe dark-surround curve in Fig. 5) in about the same range of intensities (10–1000 ml). Stiles (1929), for example, had found that the relational threshold of brightness reached its minimum of 0.6 percent at a brightness of 3.38 ml, and definitely increased at higher brightnesses. Later, however, Stiles and Crawford (1934) found that above a certain brightness the relational threshold was constant; and their brightnesses went as high as 1691 ml. Steinhardt (1936), a student of Hecht's, claimed to find a constant relational threshold at high brightnesses, though his surrounds were smaller and relatively much less bright than Stiles' (1929). His graphs show "$\Delta I/I$" to become practically constant at his highest intensities; but in this region (44–100,000 ml[6]) the theoretical (Hecht's) curves he has drawn through the data for his larger test-patch sizes (substantially larger than Lasareff's 40′ minimum) do not pay much attention to the actual datum-points. With his largest patches, "$\Delta I/I$" definitely rises again beyond 44 ml. Neither Steinhardt's work nor that of Hecht, Peskin, and Patt (1939), employing the same apparatus, affords any convincing evidence against a *rise* of "$\Delta I/I$"; and Bartlett (1942) also found a slight rise in some instances.

At most, one can say only that at high brightnesses the relational threshold perhaps attains and maintains constancy and does not ultimately increase again. Very definitely, however, it does not show the slow but endless decrease demanded by the newer visual acuity data. One can only conclude, with Lythgoe, that visual acuity continues to increase with intensity after brightness discrimination has stopped improving. Physiologists cannot forever remain contented with this "paradox."

Secondary Factors

A number of secondary factors influence visual acuity. Some of them do so simply by altering the effective level of retinal illumination or the distribution of brightness in the image, while others have their effects in more complex ways.

[6] The value of the "photon" unit of Steinhardt's abscissa is not the same for all of the curves in his Fig. 2, for two of them represent a 3-mm pupil and the rest a 2-mm pupil. Troland's photon was long since outmoded by Stiles and Crawford's (1933) discovery; and Hecht (see Hecht, 1934, pp. 776–7; Hecht and Mintz, 1939, footnote; but see Hecht, Peskin, and Patt, 1939) has returned sporadically to the expression of intensities in simple millilamberts since corrections for pupil size proved unimportant.

If the retina is not pre-adapted to the illumination of the target background, it will be as if that illumination were higher or lower, introducing either glare or low visibility, and time for adjustment must be afforded if performance is to be optimal. The exposure-time must in any case be at least an appreciable fraction of a second or the full intensity of the stimulus cannot be registered, and again it will be as if that intensity were lower. Any decrease in the contrast between target and ground (as in the substitution of gray-on-white for black-on-white targets, or even in the use of front-illuminated black-on-white targets whose black elements have some reflectance) will interfere with the production of a perceptible difference in brightness, just as if the illumination were lower. Absorption in the dioptric media of the eye is considerable, and varies with age, so that the age of the subject can never be ignored. Individual differences in the density of the intraocular filters (the yellow pigmentations of the lens and the macular region of the retina), in the diameter of foveal cones, and in the conformation of the foveal depression, are considerable and have some influence. Mysteriously, negroes have higher acuity than whites (Helson and Guilford, 1933), and there are probably consistent differences between dark-eyed and lighter-eyed whites, including partial and complete albinos.

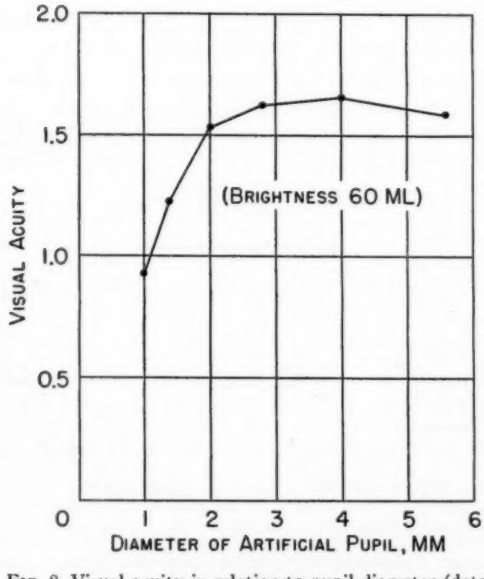


FIG. 8. Visual acuity in relation to pupil diameter (data from Cobb).

Refractive errors are of course important. Astigmatism, if uncorrected, is harmful but is not so disastrous as it is in so-called vernier acuity (see below). An *uncorrected* myope, within his range of object-distances for sharp imagery, has a larger retinal image than an emmetrope, and a hypermetrope has a smaller one. Corresponding differences in resolving power can be demonstrated.

The pupil plays a significant role. The size of the diffusion-disk from each point is determined by diffraction at the pupil, and by chromatic aberration. The supraliminal portion of the diffraction-spot decreases as the pupil enlarges, while the chromatic-aberration spot increases with the pupillary diameter. The two are in approximate balance for pupil diameters of from three to six millimeters (Hartridge, 1922). Visual acuity might be expected to be roughly constant, *ceteris paribus*, over this range; and it is: Lister (1913) found it to be about constant for grid and grille targets between 2.54 and 6.1 mm, but the thresholds of his most sharp-sighted observers did not reach minimum with pupils smaller than about four millimeters. Cobb (1915), also using a grille object, employed artificial pupils ranging from 1.0 mm to 5.6 mm. The threshold was minimal with a 4.0-mm pupil when the brightness of the target was constant and not adjusted for pupillary area to give a "constant" brightness of retinal image (Fig. 8). When this latter compensation was attempted, acuity was constant for pupils 2.0–4.0 mm in diameter and fell off but slightly with the 5.6-mm pupil; but it is now known that no such simple compensation can be made. Stiles and Crawford (1933; confirmed by Wright and Nelson, 1936; see also Stiles, 1939) found that the effective brightness of rays falls off more and more rapidly as their point of entrance is moved outward from the center of the pupil, so that with two pupils in, say, 1 : 2 ratio as to area, the retinal-image brightnesses are in nothing like 1 : 2 ratio.

Biologically, a four-millimeter pupil "should" be ideal; for this is the normal diameter (Duke-

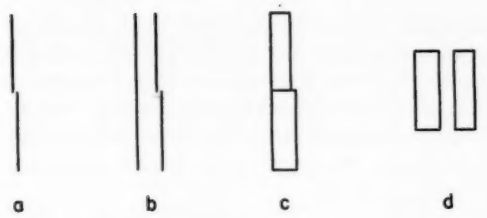


FIG. 9. Vernier acuity as a spatial perception. *a*—the customary target for coincidence settings; *b*, *c*, and *d* are explained in the text.

Elder, 1933; p. 542), and is the size to which the pupil always returns after a few minutes' adaptation to any brightness from 9 to 99 ml (Schirmer, 1894)—outside of which range the pupil size bears a fixed relation to the intensity. It will be noted that this intensity-range for the "physiological size" of the pupil embraces the values for which brightness discrimination and visual acuity are near-maximal.

The spectral character of the illumination can influence visual acuity considerably. Resolution is ordinarily higher under white light than under any monochromatic light except yellow, but it is decidedly lower under red and (especially) blue than under white or the mid-region of the spectrum (Bell, 1911; Dow, 1911; Luckiesh, 1911; Luckiesh and Moss, 1933a; Duke-Elder, 1933; Ferree and Rand, 1942). The superiority of monochromatic light from the middle of the spectrum obtains even when photometric brightnesses are equated, and the inferiority of red and blue is lessened somewhat when the considerable refractive errors which they introduce are compensated.[7] These effects are almost universally attributed to the reduction of chromatic aberration (but see Roaf, 1930), and it seems highly significant that great numbers of diurnal animals (including man—see above) have produced yellow filters somewhere in their eyes (see Walls, 1942; cf. Luckiesh and Moss, 1925). The very low value of acuity under blue illumination has, however, seemed puzzling to many, since diffraction is least with blue; and under blue light the pupil can be much smaller than under white, before its size becomes the limiting factor in visual acuity (Shlaer, Smith, and Chase, 1942).

Psychic factors exist also, though many "objective" workers on visual acuity are loth to admit it. In this category are perhaps to be placed the effects of practice and past experience, which will ordinarily manifest themselves regardless of the nature of the test-object but are notoriously great in clinical tests involving printed letters. It is generally agreed that such tests entail visual recognition rather than visual resolution, no matter how "scientific" the basis upon which they are constructed. An interesting illustration of this can be afforded by any printed list of personal names. Such a list can be placed at such a distance that an observer, unfamiliar with the list, can read all of the given names without error and yet may be unable to make out a single one of the surnames—for he has seen, somewhere, each of the given names hundreds of times for every experience he has had with any one surname.

Strictly psychical in basis (though that basis is not at all understood) are the substantial differences in visual acuity that occur with targets which subtend equal angles both as to their separate elements and their entireties, but which are unequal in actual size (and therefore, distance). Acuity may be higher for the nearer target or for the farther, depending upon circumstances (Jacobsson, 1917; von Lempicka, 1919; Freeman, 1929; Luckiesh and Moss, 1933b); and these phenomena hold for the fovea as well as for the periphery, where they were long ago discovered in the guise of the well-known "Aubert-Förster phenomenon." McFadden (1940) has found that acuity can be different for two identical targets whose perspective surroundings are such as to make them *appear* to stand at different distances.

Psychophysiological in nature are the effects of stimulation of sense-organs other than the eye in use for observation. In the crude tests employed clinically, binocular visual acuity will usually be found "exactly" equal to the acuity of the better eye used alone; but in refined laboratory experiments, binocular "summation" of acuity is claimed to take place. Reports of actual experimental demonstration are difficult

[7] The normal eye is simultaneously 0.75 diopter hypermetropic for the red and 1.25D myopic for the violet end of the spectrum; and in accommodation it seems to insist upon focusing the yellow region in the visual-cell layer even when no yellow is present in the illumination supplied.

to find, but such authorities as Duke-Elder (1933, pp. 754–5), Shaad (1934), Verhoeff (1935), and Talbot and Marshall (1941, p. 1261) either state categorically that it occurs, or at least show in context that they believe in it. Since binocular luminosities are higher than monocular (Bartley, 1941), one should perhaps expect acuity to be summated. The effects of cerebral interaction of some other kinds cannot be denied. Kravkov's original demonstration of an effect on monocular

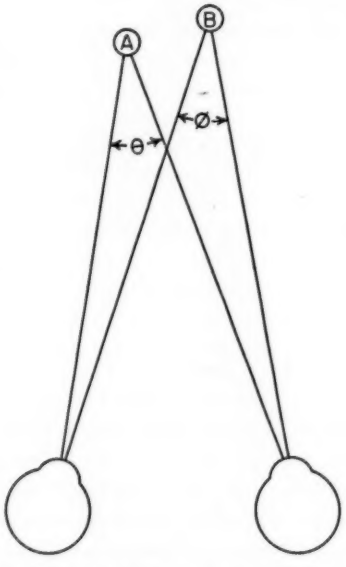


FIG. 10. Plan view of simple two-rod arrangement for the determination of stereo-acuity. The threshold $= \theta - \phi$.

acuity through the illumination of the other eye, and even through stimulation of other senses, has been repeatedly confirmed (Hartman, 1933; Kravkov, 1934; Ryan, 1940).

Travis and Martin (1934) have reported an increase in the resolution of each of two targets, simultaneously imaged on the nasal and temporal peripheries of the same retina, over the resolution of either target singly. Here, the cerebral images which are interacting are entirely in the separate cerebral hemispheres, owing to the partial decussation of the optic nerve. Such phenomena as this, the Kravkov effect, and binocular summations indicate that there are thresholds in the central visual system which are higher than those encountered by a visual impulse within the retina. That is, we do not necessarily have a sensation just because an optic nerve fiber conducts an impulse, but an ineffective impulse may be made effective by some change occurring in the higher reaches of the nervous system.

Resolution of Color-on-Color

A gentle indictment can be laid against all the investigators who have wrestled with the problems of resolution and come away from the strife proudly clasping theories to their bosoms. One and all, they seem to be color-blind: Not one has seemed to realize that for an organism with color vision and with a given capacity for spatial localization, the whole capacity for pattern discrimination is limited not by brightness discrimination alone, but by the arithmetical product (in a sense) of brightness discrimination and hue (and saturation) discrimination. Only in one investigation—that of Koffka and Harrower (1931)—has visual acuity been studied with colored targets on gray grounds of equal brightness, and gray targets on colored grounds of equal brightness. When at last someone determines visual acuities for colored targets on colored grounds of equal brightness, the advocates of brightness discrimination as the "sole" basis of resolution will be given new food for thought.

VERNIER AND STEREO-ACUITY

These kinds of visual performance do not, to most minds, involve the resolution of anything at all, though occasional investigators still presume them to have their basis in the mensuration of the visual-cell mosaic.

Just as the perception of a line and the resolution of a gap between two lines have slowly come to be seen to have a similar basis, so also vernier and stereo-acuity have been considered, since the turn of the century, to have a great deal in common—although one is a monocular capacity and the other, binocular. Vernier acuity is the accuracy of detection of a misalignment or of a variation in width. Stereo-acuity is the accuracy of detection of differences in distance from the observer. Both are spatial perceptions, largely

independent of illumination and contrast, and quite independent of the dimensions of retinal elements.

Vernier Acuity

The accuracy of setting two lines in alignment (Fig. 9*a*) is expressed by the angle which, at the eye, subtends the lines produced alongside each other. The threshold is so very small that if the lines *were* so produced, they could not possibly be resolved. Indeed, it is as small for the homologous edges of bars which, if produced, would be largely overlapped. In the standard vernier or "coincidence" situation shown in Fig. 9*a*, one does not discriminate the gap or offset between the ends of the lines, but the difference in *spatial location* of the lines. If we place a third line near the first two, it is perhaps more easily seen that spatial discrimination is involved (Fig. 9*b*); and this is further emphasized if we complete the rectangles as in Fig. 9*c*. If the widths of these rectangles are to be compared, it is not even necessary to have them in these positions; for, long before the Fig. 9*a* test-object was thought of, Volkmann (1863) found that a difference in width of two narrow rectangles could be perceived with an error of only 7–11″ arc (Fig. 9*d*).

Wülfing, in 1892, discovered the high value of acuity for the simple two-line alignment situation, and obtained thresholds of 10–12″ for transilluminated slits and for black lines on a white card. In this century, Baker and Bryan's (1912) best observers reached thresholds of about 4″ for the coincidence of either black lines on a white ground or white lines on black. "Irradiation," in the latter case, is of no consequence; and Hartridge (1922) obtained a value of 11″ even for blurred lines such as the fuzzy-edged absorption bands on spectrographic plates—which further emphasizes the distinctness of vernier acuity from ordinary visual acuity.

More recent reports give still lower values of threshold. French (1919–20) announced a figure of 0.5″, which has met with tacit skepticism. He further confirmed the fact that black-on-white and white-on-black targets give similar results, and found also that the performance holds up well under lowered illumination. Wright (1942) gives the vernier threshold as 2″; Swaine (1924), 1″. For purposes of instrument design, however, working values of 10–15″ (tracing back to the earliest experiments) are assumed in different countries.

Stratton (1900) obtained a threshold of 7″ for a target at 120 m, at a time when others were finding rather higher thresholds with nearby targets. There have been other indications that vernier acuity may improve with increasing distance, analogous to Freeman's inverted Aubert-Förster phenomenon (see above).

Stereo-Acuity

The stereo-acuity threshold is defined as the just-perceptible difference in binocular parallax

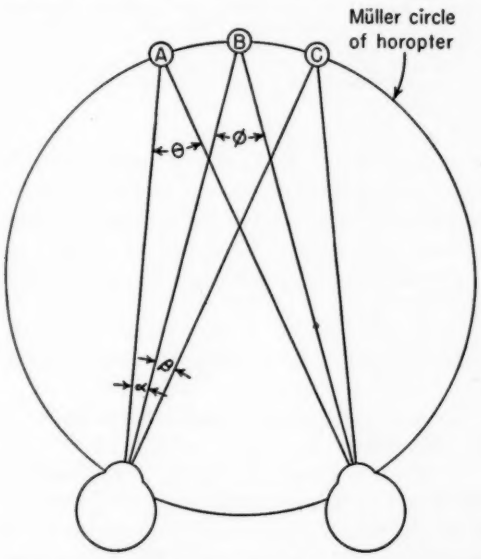


FIG. 11. Plan view of three-rod arrangement for the determination of stereo-acuity. As drawn, rod *B* is on the horopter and consequently at the null point: since θ and ϕ are then equal, *A*, *B*, and *C* will appear to be in the same plane. Further explanation in text.

of two objects or object points. In Fig. 10, the threshold is θ minus ϕ when the depth disparity of *A* and *B* is just perceptible (one does not literally perceive the difference between the angles).

Values of threshold quite comparable to vernier acuity values have been obtained at about the same times in history—it is interesting to note the similarity between the chronologies

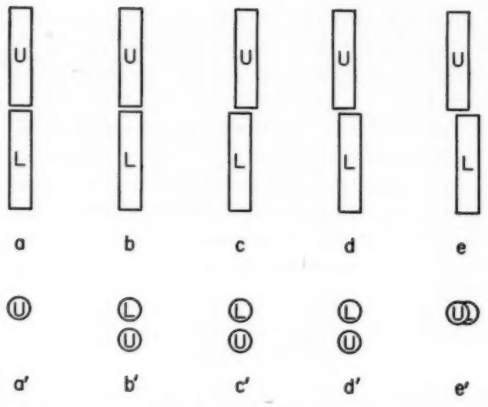


FIG. 12. Hypothetical stereo-vernier test object. *a*—without displacement—upper rod (*U*) aligned with lower rod (*L*); *b*—with *U* displaced a stereo-threshold amount toward observer's binoculus; *c* and *d*—vernieroid appearances of *b* as seen with observer's left and right eyes, respectively; *e*—with pure vernier displacement of *U*; *a'*–*e'*—corresponding plan views (*L* is completely concealed in *a'*, largely concealed in *e'*).

of these values and of athletic track and field records! In the last century, stereo-thresholds less than the subtense of a foveal cone were not attained, and were not expected owing to confusion of thought regarding the basis of the performance. Stratton (1898) was the first to break this mental deadlock, with a threshold of 24″ obtained by an indirect method. Around 1900 it was considered remarkable when Pulfrich reported a subject whose threshold was 10″; but Bourdon (1900) reached 5″, and Breton (1920) 4″. In the last decades, values of two seconds or a bit less have been regularly obtained. One-third of all of Howard's (1919) observers performed this well, as did Woodburne (1934) in a careful experiment in which the angular subtenses of the target elements were kept constant for all positions. As with vernier acuity, however, a working value of about 12″ is used in American instrumental practice; and the allowable error in the Howard-Dolman test of the Army Air Corps is 10.3″.

Relationship of Vernier and Stereo-Acuities

The stereo-situation gives the appearance of being much more complex than the vernier one, owing in large part to the existence, in binocular vision, of the horopter—the geometrical figure containing all spatial points which are seen without physiological diplopia together with the fixation point of the moment. In the usual laboratory stereo-test-object, either two rods are used (Fig. 10), or more commonly three equally-spaced wires or rods are used, with the middle one made movable in and out of the plane of the other two. The stereo-threshold is then the difference between θ and ϕ (Fig. 11) when *B* is just perceptibly behind or in front of the plane of *A* and *C*. It will be minimal when all three rods are close to the horopter, but it will then be subject to a sort of constant error since the "null point" is on the horopter itself (Fruböse and Jaensch, 1923). That is, in the situation shown in Fig. 11, all three rods would appear to lie in one plane and *B* would have to be placed outside the Müller circle, or inside of it by an equal amount, to appear disparate in distance compared with *A* and *C*.

Now in this situation, $\theta - \phi$ is the same for either eye and is the same for one eye as for two. This angular disparity is the only objective information available to the brain, upon which to base a stereo-displacement percept. One may then wonder why the response should be more

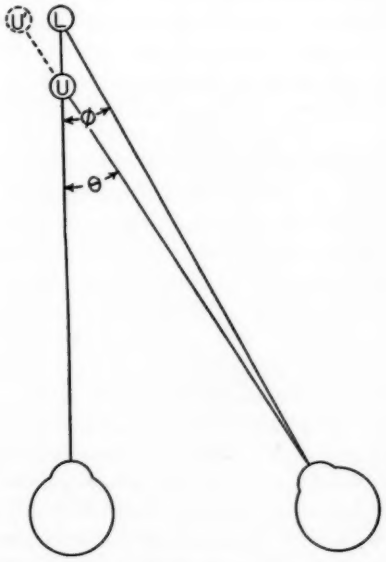


FIG. 13. The relationship of vernier and stereo-acuities as brought out with the stereo-vernier test object (see text).

precise than two-line resolution; and especially one may wonder why the depth threshold becomes so enormous when one eye is covered. Why should not monocular distance judgment be just as good as binocular, considering that binocular vision has nothing more to go on than two identical pieces of monocular information? When θ and ϕ are made equal—as they automatically are when all three rods are on the Müller circle as in Fig. 11—the binocular basis for the perception of depth-disparity disappears as we have seen; but one can say that this is because the monocular subtenses α and β have also automatically become equal.

The basis of stereo-acuity and its closeness to the basis of vernier performance can perhaps be grasped most readily from a consideration of a hypothetical apparatus which might be called the "stereo-vernier test object." This would be formed by two rods, one above the other, with one—say, the upper—made movable toward and away from the observer and also laterally, in the same frontal plane as the other rod (Fig. 12).

Suppose that the plane of displacement of this target contains the left visual axis, and the stereo-displacement is made of threshold value

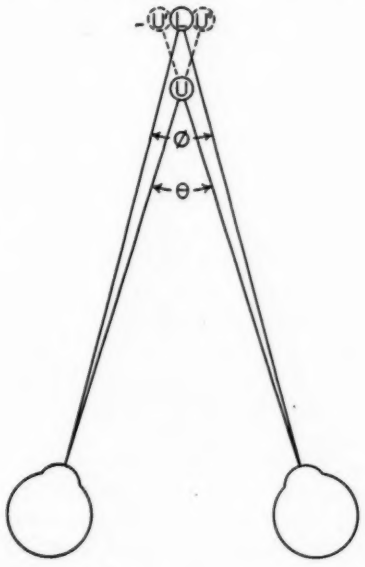


FIG. 14. The relationship of vernier and stereo-acuities as brought out with the stereo-vernier test object (see text).

FIG. 15. Illustrates the fact that the disparity between the left- and right-eyed widths of a three-dimensional object does not exist on a single retina and hence does not need actually to be resolved, in the sense of monocular parallel-line resolution. The limit of binocular stereo-acuity is consequently not set by the factors which determine parallel-line resolving power.

(Fig. 13). If the right eye is now covered, the left eye cannot detect whether the rods are in the position of Fig. 12*a* or that of Fig. 12*b*. With both eyes in use, the stereo-threshold is $\theta - \phi$ as usual; but all of this difference-angle exists for the right eye alone. That eye—if the left one be covered—has no way of distinguishing whether U has indeed been displaced forward, or in a purely vernier fashion into the position U' (cf. Fig. 12*e*). It is clear, however, that if the disparity of U and L is perceptible stereoscopically, an actual displacement $L \rightarrow U'$ should be detectible monocularly, as a vernier displacement; and the displacement $L \rightarrow U$ should be perceptible monocularly as a *vernieroid* displacement of the same magnitude. Angularly, $L \rightarrow U$ and $L \rightarrow U'$ are entirely equivalent, for both are measured as $\theta - \phi$.

Consider now, however, the arrangement in Fig. 14, where the target is centered and the stereo-displacement is made along the median visual axis—that of the observer's "binoculus" or hypothetical cyclopean eye. If the displacement $L \rightarrow U$ is just liminal, the appearance binocularly can only be that of Fig. 12*b*. Since the equivalent vernier displacements $L \rightarrow U'$ and $L \rightarrow U''$ here are each only half of the $L \rightarrow U'$ of Fig. 13, neither may be perceptible monocularly—the observer with either eye covered will not have the clear-cut appearances of Fig. 12*c* and *d*. Only the displacement in depth, "straight on" as in Fig. 12*b*, will be discernible; but this is created from two vernieroid displacements which, separately, should be subliminal since they represent the just-liminal $L \rightarrow U'$ of Fig. 13, halved between the two eyes instead of being placed entirely in one eye. The essence of this

idea is far from new, for Stratton (1898) was perhaps the first to point out that stereoscopic acuity arises from differences which may not of themselves be perceptible, but which arouse a stereoscopic effect. If, in the Fig. 14 situation, when one eye is covered, a vernieroid displacement *is* perceptible (cf. Fig. 12*c*, *d*), it cannot possibly be perceived as a displacement in *depth* unless $L \rightarrow U$ is made far greater. Binocular stereo-acuity is thus on the same objective basis as monocular vernier acuity, but in stereo-acuity there is an additional subjective factor: the appreciation of a displacement, which has a vernieroid aspect, *as a displacement in depth.* Only actual experimental use of the stereo-vernier target will show whether, with the same observer at the same time, stero-acuity in angular terms can ever be as good as, or better than, vernier acuity.

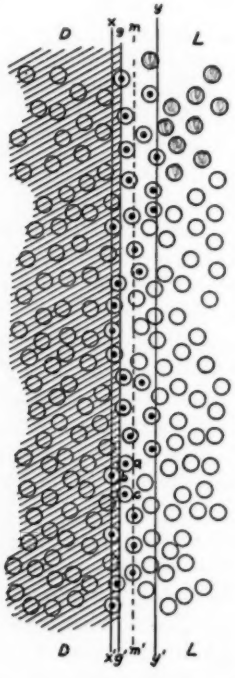


FIG. 16. Illustrates Weymouth's theory of vernier and stereo-acuity (after Andersen and Weymouth). The diagram shows retinal conditions at the margin of a stimulated area. *D–D* is in darkness, *L–L* illuminated. The geometrical margin of the image is *g–g'*. The cones are shown as circles (×1500; traced from a photomicrograph). Cones *a*, *b*, and *c* have local signs whose "center of gravity" lies amidst them, tending to pull *b* to the right. This action, among all of the cones cut by *g–g'*, smooths the percept of the contour despite the raggedness of the line of cones concerned. Furthermore, normal nystagmus shifts *g–g'* back and forth between the extreme positions *x–x'* and *y–y'*, so that *m–m'* represents the center of gravity of all the points stimulated, and is hence the local sign of the percept to which they give rise. The localization of this percept is independent of such considerations as the size of one cone.

Explanations

In the commonplace stereoscopic situation one is, essentially, sensing simultaneously two widths. Each of these widths (the left- and right-eyed aspects of a space, or of a facet of an object) is being presented to a separate eye. It is important to note that the excess of one width over the other, thought of as a left-over strip along one side of the "fused" image (Fig. 15), is not having to be resolved in a single image. It is "resolved" in the fusion image, where nothing can interfere with the recording of the true spatial positions of the various contours in each monocular aspect of the object—no matter how nearly "coincident" their images may be—within the same limits of spatial localization which govern vernier performances.

Even so, the small magnitudes of the stereo- and vernier thresholds have disturbed many an investigator, unable to shut out of his mind the idea that the "grain" of the visual-cell mosaic ought somehow to limit these capacities. The continuousness of visual space—its total lack of resemblance to a magnified half-tone reproduction—has seemed a great puzzle. No less puzzling is the continuousness of the movement of a moving object, for some seem to think that a moving object should always look as if it were on the far side of a picket fence.

One apparently "sees" even with the blind bits of retinal area between the visual cells, and one can perceive object-points as lying in spatial positions intermediate between those on the projected axes of adjacent cones. How can such things be?

Hering accounted for vernier acuity by proposing that one of the two contours just out of alignment is imaged just within one row of cones and the other along the edges of the cones in the next row, having a different "local sign" from those of the first row. Modern knowledge of the absence of any such precise alignment of the

cones, the near-equality of vernier acuity with targets oriented in various positions other than the vertical, and the shadowy character of retinal "images," has combined to make Hering's view seem altogether too atomistic for the twentieth century. But no better theory was forthcoming until the publications of Weymouth and his associates (Andersen and Weymouth, 1923; Weymouth, Andersen, and Averill, 1923; Averill and Weymouth, 1925; Weymouth, Hines, Acres, Raaf, and Wheeler, 1928).

Weymouth *et al.* made deliberate use of the irregularities of the cone mosaic, and of the continuous and involuntary oscillations of the eyeballs which others had always supposed to *detract* from resolution performances. Weymouth's theory grew out of Hering's doctrine of retinal local signs. Each cone was supposed to have such a sign, meaning that if a particular cone were stimulated, the resulting percept of a point of light would be projected along a unique direction into visual space. Weymouth proposed that when two adjacent cones were stimulated, the local sign of each would "partake of the nature of the other," by which was apparently meant that a single point of light could then be seen in an intermediate position in space, leaning toward a position on the produced axis of one or the other of the cones in accordance with the relative stimulation of them. This procedure would work out to create the percept of smooth line even when a very wiggly line of cones was being stimulated (Fig. 16).

The simultaneous stimulation of cones just on either side of an edge of an image would help to place the contour perceptually intermediate between them. The "mean local sign" of all the cones along the edge of the image would yield a spatial projection of the perceptual contour in quite exact correspondence with the real position of the objective contour.

Obviously, the longer the contour the better this mechanism should work; and Weymouth made much of the fact that increasing the length of stimulus lines or rods, up to at least the subtense of the macula, had long been known to promote vernier and stereo-acuity. Weymouth also believed in a temporal integration of local signs, in that the physiological nystagmic fluttering of the eyeball, far from interfering with resolution, would sweep the image contours over so many irregular lines of cones that the grand average mean retinal local sign would produce a most precise localization of the objective contours in space. By way of analogy, Weymouth made the questionable statement that the average of a great number of careless measurements is more valid than the average of a few careful measurements. More to the point, in explanation of how it is that we can measure space to a fraction of a cone-diameter with a retinal ruler whose calibration is in whole cone-diameters, is Weymouth's reminder that the average of a number of measurements made in whole inches can be accurate to any desired small fraction of an inch. Given long enough stimulus-lines, the foveal cones apparently might as well be far larger than they are, and localization acuities would remain unchanged.

Weymouth *et al.* thought they had proven the necessity of the physiological nystagmus for the best performance, when they reduced the presentation-time of the stimulus to the point where insufficient time was afforded for any eye movement. With presentations of less than 0.1 second, the threshold obtained with lengthy presentations was multiplied by five. But these workers obtained such surprisingly *high* thresholds at best that the significance of some of their procedural modifications is in doubt.

Langlands (1929) found the stereo-acuity threshold to be higher under brief electric spark illumination, and lower and lower with increasing presentation-times. From 10^{-7} second to 100 milliseconds there was little change. Acuity rose sharply from 100 to 500 milliseconds, then more slowly to four seconds, and more slowly still for several seconds thereafter. Considering that the period of the continuous fine oscillations of the eye, covering an angle of about 2′, is from 10 to 20 milliseconds (Adler and Fliegelman, 1934), it seems impossible that acuity should be subject to influence over such a wide range of presentation-times unless retinal adaptation to intensity is involved (see Eguchi, 1932). Critical examination of Langlands' technique shows that he, like others who have tried to "stop" eye movement by shortening the presentation-time, failed utterly to provide an apparatus with which the whole situation effective upon the retina was

really the same except for the variation of presentation-time alone. If we are ever to know whether physiological nystagmus helps or hinders vernier resolution or any other kind of resolution, someone must have the courage actually to immobilize the eyeball.

A vital weakness of Weymouth's theory is its restriction of all the important processes to the eyeball itself, and its abuse of the local sign doctrine in its assumption that the local signs are not fixed but are subject to change by neighboring processes. No theory of localization which ignores the visual cortex and way-stations can be successful.

The central visual apparatus has received plenty of attention—perhaps too much—in the recent theory of Talbot and Marshall (1941; Marshall and Talbot, 1942). On the basis of their work on the cortical projection of retinal points in cats and monkeys, these authors have erected a ponderous machinery with which they can explain almost everything that happens in vision. It is based upon supposed anatomical relationships and to an even greater extent upon such neurophysiological phenomena as recovery cycles, multiplication of paths, reciprocal synaptic overlap, facilitation, channeling, funneling, and peaking.

These matters are most difficult for all but a chosen few—not including the writer—to comprehend fully. One tangible experimental fact is that two minutes of arc of the Rhesus monkey's fovea (=*ca.* 9μ) project to one linear millimeter of the visual cortex. Talbot and Marshall believe that the effective ratio of cortical ganglion cells to foveal cones is at least 600 : 1, which could account for vernier and stereo-thresholds very, very much smaller even than those actually observed—whose smallness is already so frightening. The situation as they expound it is diagrammed, no doubt with great oversimplification, in Fig. 17.

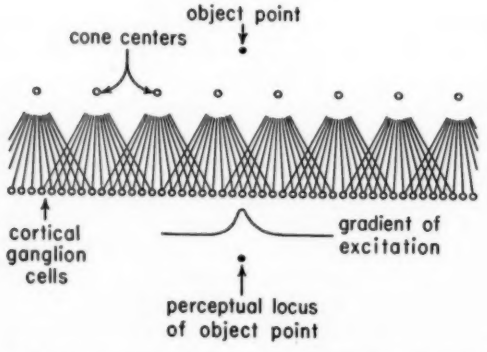


FIG. 17. Illustrates Talbot and Marshall's explanation of the precision of spatial localization, as the basis of vernier and stereo-acuities. Reciprocal overlap in the central visual pathway, together with neurophysiological phenomena, creates a gradient of cortical excitation whose peak is the locus of the percept of the object point. A great number of such loci can lie along the line representing the center-to-center distance between adjacent retinal cones.

If differential stimulation of adjacent visual cells can place the peak of excitation in any one of the cortical loci opposite the space between the visual-cell centerlines, then we have only to pull Hering's local signs up onto the *cortical* visual cells (where they have probably always belonged) in order to account for the precision of spatial localization. Talbot and Marshall show great respect for some aspects of Weymouth's theory, for which they have provided this central-nervous basis. Even the physiological nystagmus is grist to their mill, for they credit it with producing the *sudden* changes of illumination of individual cones which, they feel, are just as important for brightness discrimination as the *amount* of change. Also, it augments facilitation and thus sharpens the neural "focus" of the cerebral image; for, the period of the finest bulbar oscillations is nicely related to the temporal aspects of the cycle of recovery from conduction in nerve fibers. Talbot and Marshall say nothing about moving objects; but their mechanism patently accounts for the continuousness of visual movement since the cortical "grain" is so much finer than the retinal.

But, now, does any such tremendous morpho-physiological subdivision of retinal grain actually occur in the central visual apparatus? It seems a little silly on the face of things; for in all retinae, and particularly in nocturnal ones, there is a great reduction of path number from visual cells to optic nerve fibers, which is purposed to increase over-all retinal sensitivity. It would seem that this effect would be undone by any corresponding multiplication of paths occurring between retina and cortex. Reduction followed by multiplication might be intended to create a slender optic nerve and a small blindspot; but

fishes perform the reduction in the retina in the absence of any opportunity for a counteractive multiplication in their acorticate brains. And, fishes see ϕ-movements just as they see real ones, indicating that they see real movements as continuous (see Walls, 1942).

In the primary visual pathway from retina to cortex there is only one way-station—the lateral geniculate nucleus—in which optic nerve fibers synapse with another set of neurons extending through the optic radiations to the visual cortex. The monkey brain is not a human one, and there are some surprising differences between their visual systems. No one knows the situation in the human (Hines, 1942); but in the monkey's lateral geniculate, Glees and Clark (1941) found that each optic nerve fiber gives off five or six terminal branchlets, each ending in a button applied to the body of a geniculate ganglion cell. Each of these cells (whose axons run to the cortex) receives only one such button, so that there is no reciprocal overlap.[8] Thus not more than six fibers in the geniculo-cortical tract correspond to one optic nerve fiber, and (in the foveal projection) to one cone.

This is far from the *ca.* 600 : 1 ratio proposed by Talbot and Marshall; and in fact there may be no effective "multiplication of paths" at all, for in stereopsis several cortical cells may be required to deal with the localization of one object-point. But if one *must* think of vernier performance as a resolution of something, it may be noted that the Rhesus monkey's vernier threshold might be expected to be one-sixth of his threshold of resolution for parallel lines. In man, the ratio is more nearly one-twelfth—two seconds or less as against twenty-four seconds (though Lythgoe's work indicates that the latter figure may sometime be bettered). Future work may indicate an anatomical basis for such a ratio in man; but as previously pointed out, it seems quite unnecessary to consider vernier and stereo-performances as bearing any relation to ordinary visual acuity. Without reciprocal overlap in the geniculate, it is hard to see how there can be any more cortical local signs than there are retinal receptors; but Clark has shown that short associational fibers can spread the excitation from one cortical cell-body to others as much as five millimeters distant. This provision for "overlap" may be the real basis of spatial localization.

[8] Clark (1942) notes that Glees found reciprocal overlap to be characteristic of the lateral geniculate of the cat, which is a nocturnal animal. He suggests, very reasonably, that the overlap is the basis of high sensitivity to light rather than of high acuity. He points out that more geniculate cells project to one square millimeter of cortex in the nocturnal, low-resolving rat than in the diurnal, sharp-sighted monkey.

BIBLIOGRAPHY

F. H. Adler, "Problems in the physiology of visual acuity," Arch. Ophthal. **11**, 6–19 (1934).

F. H. Adler and M. Fliegelman, "Influence of fixation on the visual acuity," Arch. Ophthal. **12**, 475–483 (1934).

E. E. Anderson and F. W. Weymouth, "Visual perception and the retinal mosaic. I. Retinal mean local sign—an explanation of the fineness of binocular perception of distance," Am. J. Physiol. **64**, 561–594 (1923).

H. L. Averill and F. W. Weymouth, "Visual perception and the retinal mosaic. II. The influence of eye-movements on the displacement threshold," J. Comp. Psychol. **5**, 147–176 (1925).

T. Y. Baker and G. B. Bryan, "Errors of observation," in: *The Proceedings of the Optical Convention 1912*, **2**, 253–265 (Hodder and Stoughton, London, 1912).

E. E. Barnard, "A micrometrical determination of the dimensions of the planets and satellites of the solar system made with the 36-inch refractor of the Lick Observatory," Pop. Astron. **5**, 285–302 (1897–8).

N. R. Bartlett, "The discrimination of two simultaneously presented brightnesses," J. Exper. Psychol. **31**, 380–392 (1942).

S. H. Bartley, *Vision* (D. Van Nostrand Company, Inc., New York, 1941).

L. Bell, "Chromatic aberration and visual acuity," Elec. World **57**, 1163–1166, 1569 (1911).

B. Bourdon, "L'acuité stéréoscopique," Rev. Phil. de la France et de l'Étrang. **49**, 74–78 (1900).

W. K. D. Breton, "Rangefinders and the standard of fitness required in rangetakers," J. Roy. Nav. Med. Serv. **6**, 288–307 (1920).

W. Le Gros Clark, "The visual centres of the brain and their connexions," Physiol. Rev. **22**, 205–232 (1942).

P. W. Cobb, "The influence of pupillary diameter on visual acuity," Am. J. Physiol. **36**, 335–346 (1915).

P. W. Cobb and L. R. Geissler, "The effect on foveal vision of bright surroundings," Psychol. Rev. **20**, 425–447 (1913).

P. W. Cobb and F. K. Moss, "The four variables of the visual threshold," J. Frank. Inst. **205**, 831–847 (1928).

J. S. Dow, "Visual acuity and light of different colors," Elec. World **58**, 955–956 (1911).

W. S. Duke-Elder, *Text-Book of Ophthalmology* (C. V. Mosby Co., St. Louis, 1933), Vol. 1.

H. Eguchi, "Course of improvement of visual acuity during light adaptation," *Proceedings of the International Illumination Congress 1931* (International Illumination Congress, London, 1931, 1932), Vol. 1.

C. Fabry, "Vision in optical instruments," Proc. Phys. Soc. **48**, 747–762 (1936).

C. E. Ferree and G. Rand, "Color and composition of light in relation to the blackout," J. Aviation Med. **13**, 193–215 (1942).

C. E. Ferree and G. Rand, "Wartime and blackout lighting in relation to the eye," Illum. Eng. **37**, 579–595 (1942).

E. Freeman, "Anomalies of peripheral visual acuity," J. Exper. Psychol. **12**, 324–340 (1929).

E. Freeman, "What does a test of visual acuity measure?" Arch. Ophthal. **2**, 48–56 (1929).

J. W. French, "The unaided eye," Trans. Opt. Soc. **20**, 209–236; **21**, 1–33; **21**, 127–156 (1919–20).

A. Fruböse and P. A. Jaensch, "Der Einfluss verschiedener Faktoren auf die Tiefensehschärfe," Zeits. f. Biol. **78**, 119–132 (1923).

G. A. Fry and P. W. Cobb, "Visual discrimination of two parallel bright bars in a dark field," Am. J. Psychol. **49**, 76–81 (1937).

P. Glees and W. E. Le Gros Clark, "The termination of optic fibres in the lateral geniculate body of the monkey," J. Anat. **75**, 295–308 (1941).

G. W. Hartman, "The increase of visual acuity in one eye through illumination of the other," J. Exper. Psychol. **16**, 383–392 (1933).

H. Hartridge, "Visual acuity and the resolving power of the eye," J. Physiol. **57**, 52–67 (1922).

H. Hartridge, "Physiological limits to the accuracy of visual observation and measurement," Phil. Mag. **46**, 49–79 (1923).

S. Hecht, "A quantitative basis for visual acuity and intensity discrimination," Skand. Arch. f. Physiol. **49**, 146 (1926).

S. Hecht, "A quantitative basis for the relation between visual acuity and illumination," Proc. Nat. Acad. Sci. **13**, 569–574 (1927).

S. Hecht, "The relation between visual acuity and illumination," J. Gen. Physiol. **11**, 255–281 (1928).

S. Hecht, "Visual acuity and illumination," Arch. Ophthal. **57**, 564–573 (1928).

S. Hecht, "Entgegnung an Best," Naturwiss. **18**, 237–239 (and see *ibid.*, 233–237) (1930).

S. Hecht, "Vision: II. The nature of the photoreceptor process," (Chapter 14, pp. 704–828, in: *A Handbook of General Experimental Psychology;* Clark University Press, Worcester, 1934).

S. Hecht, "Intensity discrimination," in: *Cold Spring Harbor Symposia on Quantitative Biology*, Vol. 3, pp. 230–236 (1935).

S. Hecht, "A theory of visual intensity discrimination," J. Gen. Physiol. **18**, 767–789 (1935).

S. Hecht and E. U. Mintz, "The visibility of single lines at various illuminations and the retinal basis of visual resolution," J. Gen. Physiol. **22**, 593–612 (1939).

S. Hecht, J. C. Peskin, and M. Patt, "Intensity discrimination in the human eye. II. The relation between $\Delta I/I$ and intensity for different parts of the spectrum," J. Gen. Physiol. **22**, 7–19 (1939).

M. Heinz and F. Lippay, "Über der Beziehungen zwischen der Unterschiedsempfindlichkeit und der Zahl der erregten Sinneselemente," Arch. f. d. ges. Physiol. **218**, 437–447 (1928).

H. Helmholtz, *Handbuch der physiologischen Optik* (Leopold Voss, Leipzig, 1867).

H. Helson and J. P. Guilford, "The relation of visual sensitivity to the amount of retinal pigmentation," J. Gen. Psychol. **9**, 58–76 (1933).

M. Hines, "Recent contributions to localization of vision in the central nervous system," Arch. Ophthal. **28**, 913–937 (1942).

F. B. Hofmann, "Die Lehre vom Raumsinn, I. Teil," in: *Handbuch der gesamten Augenheilkunde* (ed. Graefe-Saemisch) (Leipzig: Engelmann, Leipzig, 1919), second edition, Vol. 3, Chapter 13, pp. 1–213.

H. J. Howard, "A test for the judgment of distance," Am. J. Ophthal. **2**, 656–675 (1919).

M. Jacobsson, "Über die Erkennbarkeit optischer Figuren bei gleichem Netzhautbild und verschiedener scheinbarer Grösse," Zeits. f. Psychol. u. Physiol. d. Sinnesorgane I (Zeits. f. Psychol.) **77**, 1–91 (1917).

K. Koffka and M. R. Harrower, "Colour and organization. Part II," Psychol. Forsch. **15**, 193–275 (1931).

A. König, "Die Abhängigkeit der Sehschärfe von der Beleuchtungsintensität," Sitz. preuss. Akad. Wiss. 559–575 (1897).

A. König and E. Brodhun, "Experimentelle Untersuchungen über die psychophysische Fundamentalformel in Bezug auf den Gesichtssinn," Sitz. preuss. Akad. Wiss. 917–931 (1888); 641–644 (1889).

S. V. Kravkov, "Changes in visual acuity in one eye under the influence of the illumination of the other or of acoustic stimuli," J. Exper. Psychol. **17**, 805–812 (1934).

N. M. S. Langlands, "Experiments on binocular vision," *Medical Research Council* (British), *Special Report Series No. 133 (Reports of the Committee Upon the Physiology of Vision, No. IV)* (H. M. Stationery Office, London, 1929).

C. Lapicque, "Image rétinienne d'un point éloigné pour différentes grandeurs de la pupille," Comptes Rendus **202**, 656–659 (1936).

P. Lasareff, "Studien über das Weber-Fechner'sche Gesetz. Einfluss der Grösse des Gesichtsfeldes auf den Schwellenwert der Gesichtsempfindung," Arch. f. d. ges. Physiol. **142**, 235–240 (1911).

Y. Le Grand, "Sur la mesure de l'acuité visuelle au moyen de franges d'interférence," Comptes Rendus **200**, 490–491 (1935).

W. von Lempicka, "Räumliche Farbenmischung auf der Netzhaut," Zeits. f. Psychol. u. Physiol. d. Sinnesorgane I (Zeits. f. Sinnesphysiol.) **50**, 217–251 (1919).

J. J. Lister, "On the limit to defining-power, in vision with the unassisted eye, the telescope, and the microscope," J. Roy. Micros. Soc. **33**, 34–55 (1913; dated 1842–1843).

P. Lowell, "Experiment on the visibility of fine lines in its bearing on the breadth of the 'canals' of Mars," Bull. Lowell Obs. **1**, No. 2 (1903).

M. Luckiesh, "Monochromatic light and visual acuity," Elec. World **58**, 450–452 (1911).

M. Luckiesh, "The dependence of visual acuity on the wave-length of light," Elec. World **58**, 1252–1254 (1911).

M. Luckiesh, "Visual acuity and light of different colors," Elec. World **58**, 1255 (1911).

M. Luckiesh and F. K. Moss, "The effect on visual acuity of shortening the spectrum at the blue end," J. Opt. Soc. Am. **10**, 275–281 (1925).

M. Luckiesh and F. K. Moss, "Visual acuity and sodium-vapor light," J. Frank. Inst. **215**, 401–410 (1933a).

M. Luckiesh and F. K. Moss, "The dependency of visual acuity upon stimulus-distance," J. Opt. Soc. Am. **23**, 25–29 (1933b).

R. J. Lythgoe, "The measurement of visual acuity," *Medical Research Council* (British), *Special Report Series No. 173 (Reports of the Committee Upon the Physiology of Vision, No. X)* (H. M. Stationery Office, London, 1932).

H. B. McFadden, *Visual Acuity* (Optometric Extension Program, Duncan, Oklahoma; and as thesis, Ohio State University; 1940).

W. H. Marshall and S. A. Talbot, "Recent evidence for neural mechanisms in vision leading to a general theory of sensory acuity," in: *Biological Symposia*, Vol. 7 (*Visual Mechanisms*), 117–164 (The Jaques Cattell Press, Lancaster, 1942).

P. Moon, *The Scientific Basis of Illuminating Engineering* (McGraw-Hill Book Company, Inc., New York, 1936).

W. H. Pickering, "Report on Mars, No. 11," Pop. Astron. **23**, 569–588 (1915).

A. Pütter, "Studien zur Theorie der Reizvorgänge," Arch. f. d. ges. Physiol. **171**, 201–261 (1918).

H. E. Roaf, "Visual acuity in light of different colours," Proc. Roy. Soc. London **B106**, 276–292 (1930).

C. O. Roelofs, "Le minimum separabile et la plus petite largeur de sensation," Arch. Néerl. de Physiol. de l'Homme et des Anim. **2**, 199–216 (1918).

T. A. Ryan, "Interrelations of the sensory systems in perception," Psychol. Bull. **37**, 659–698 (1940).

O. Schirmer, "Untersuchungen zur Physiologie der Pupillenweite," Arch. f. Ophth. **40**, Section 5, 8–21 (1894).
D. J. Shaad, "Binocular brightness summation in dark adaptation," Arch. Ophthal. **12**, 704–708 (1934).
S. Shlaer, E. L. Smith, and A. M. Chase, "Visual acuity and illumination in different spectral regions," J. Gen. Physiol. **25**, 553–569 (1942).
J. Steinhardt, "Intensity discrimination in the human eye: I. The relation of $\Delta I/I$ to intensity," J. Gen. Physiol. **20**, 185–209 (1936).
W. S. Stiles, "The effect of glare on the brightness difference threshold," Proc. Roy. Soc. London **B104**, 322–351 (1929).
W. S. Stiles, "The directional sensitivity of the retina and the spectral sensitivities of the rods and cones," Proc. Roy. Soc. London **B127**, 64–105 (1939).
W. S. Stiles and B. H. Crawford, "The luminous efficiency of rays entering the eye pupil at different points," Proc. Roy. Soc. London **B112**, 428–450 (1933).
W. S. Stiles and B. H. Crawford, "The liminal brightness increment for white light for different conditions of the foveal and parafoveal retina," Proc. Roy. Soc. London **B116**, 55–102 (1934).
G. M. Stratton, "A mirror pseudoscope and the limit of visible depth," Psychol. Rev. **5**, 632–638 (1898).
G. M. Stratton, "A new determination of the minimum visible and its bearing on localization and binocular depth," Psychol. Rev. **7**, 429–435 (1900).
W. Swaine, *Ophthalmo-Optical Manual* (The Hatton Press, Ltd., London, 1924).
S. A. Talbot and W. H. Marshall, "Physiological studies on neural mechanisms of visual localization and discrimination," Am. J. Ophthal. **24**, 1255–1264 (1941).
L. E. Travis and R. Martin, "A study of retinal summation," J. Exper. Psychol. **17**, 773–786 (1934).
L. T. Troland, "An analysis of the literature concerning the dependency of visual functions upon illumination intensity. A report prepared for the committee on industrial lighting of the National Research Council," Trans. Illum. Eng. Soc. **26**, 107–196 (1931).
F. H. Verhoeff, "A new theory of binocular vision," Arch. Ophthal. **13**, 151–175 (1935).
A. W. Volkmann, *Physiologische Untersuchungen im Gebiete der Optik* (Breitkopf und Härtel, Leipzig, 1863).
G. L. Walls, *The Vertebrate Eye* (Cranbrook Institute of Science, Bloomfield Hills, Michigan, 1942).
F. W. Weymouth, E. E. Andersen, and H. L. Averill, "Retinal mean local sign; a new view of the relation of the retinal mosaic to visual perception," Am. J. Physiol. **63**, 410–411 (1923).
F. W. Weymouth, D. C. Hines, L. H. Acres, J. E. Raaf, and M. C. Wheeler, "Visual acuity within the area centralis and its relation to eye movements and fixation," Am. J. Ophthal. **11**, 947–960 (1928).
W. W. Wilcox, "An interpretation of the relation between visual acuity and light intensity," J. Gen. Psychol. **15**, 405–435 (1936).
W. W. Wilcox and D. McL. Purdy, "Visual acuity and its physiological basis," Brit. J. Psychol. **23**, 233–261 (1933).
L. S. Woodburne, "The effect of a constant visual angle upon binocular discrimination of depth differences," Am. J. Psychol. **46**, 273–286 (1934).
W. D. Wright, "The functions and performance of the eye," J. Sci. Inst. **19**, 161–165 (1942).
W. D. Wright and J. H. Nelson, "The relation between the apparent intensity of a beam of light and the angle at which the beam strikes the retina," Proc. Phys. Soc. **48**, 401–405 (1936).
E. A. Wülfing, "Ueber den kleinsten Gesichtswinkel," Zeits. f. Biol. **29**, 199–202 (1892).

Letter to Editor

IN a recent paper[1] there was an error of statement which should be set right. On page 44 the statement was made that Eq. (12) was used in solving Eqs. (10) and (11). As a matter of fact Eq. (12) was not used and was not necessary. All that was needed, and all that was used, was the introduction of the boundary conditions laid down in Fig. 1,

$$x=0, \quad a=a_0, \quad i=i_0,$$
$$x=t, \quad a=a_t, \quad b=b_t=r(a_t+i_t\cos\zeta).$$

With this statement the paper is correct throughout, or at any rate presents correctly the intentions of the author. In the general case of an absorbing and scattering medium Eq. (12) is not true, because the energies of the incident and emergent streams of radiation are not equal if there is an uncompensated loss by absorption. In the case of zero absorption Eq. (12) is true. The above error arose from the fact that several years elapsed between the working out of the paper and its hammering into shape for publication.

E. O. HULBURT

July 15, 1943

[1] E. O. Hulburt, J. Opt. Soc. Am. **33**, 42–45 (1943).

Some Aspects of the Eye as an Image-Forming Mechanism*

KENNETH N. OGLE
Research Division, Dartmouth Eye Institute, Hanover, New Hampshire

THIS paper is concerned with the problem of the size of the image upon the retina of an eye, the refractive error of which is corrected by an ophthalmic lens, and, in connection with that problem, with the more general question of the origin of the spherical refractive errors in the eye. This first problem deals with the relative size of the retinal image of an eye corrected for refractive anomalies as compared with that of the normal eye. Because the size of the retinal image would have a bearing upon visual acuity, the question has frequently been raised whether the eye corrected by a lens would have a higher or a lower visual acuity than that of the normal eye. As computation shows, this difference in visual acuity would probably not be perceptible except in eyes in which the crystalline lens has been removed surgically (aphakia) because of cataractous conditions, where an increase of nearly 25 percent might be expected.

Until the last few years this discussion would have been considered more or less academic because small differences were considered of no consequence either monocularly or binocularly. In aphakic eyes where larger differences would exist, there was a definite binocular problem. If, from the more general point of view of binocular vision, differences in the relative size of the ocular images in the two eyes which are smaller than those occurring in aphakia are considered unessential to efficient binocular vision, it is irrelevant whether the correction of refractive anomalies modifies the size of the retinal image or not. If, however, small differences in the relative size of the ocular images are considered important, as they are now believed to be, the correction of these refractive anomalies must be treated with strict regard to those conditions affecting the size of the retinal image.

Under optically ideal conditions an unaccommodated eye is said to be normal, i.e., emmetropic when parallel rays of light converge upon the retina to form the smallest circle of diffusion; that is, when the image of a distant object is sharply defined upon the retina. (See Fig. 1.)

According to elementary Gauss paraxial-ray theory the magnification M_0 of the retinal image of an object at the finite distance u from the eye will be given by $M_0=i_0/O=-1/(1-\phi_0 u)$, where ϕ_0 and i_0 are the power and the image size of the emmetropic eye, respectively, and O is the size of the object.

One thinks of the normal eye as becoming ametropic, that is, subject to a refractive error, in two ways: (a) a change in the axial length of the eyeball, which is designated as "axial ametropia," and (b) a change in the power of the dioptric system of the eye through changes in curvature of the corneal or the lenticular surfaces, or through change in separation of cornea and lens, or through changes in the indices of refraction of the several ocular substances. This refractive error is designated as a "refractive ametropia" (either curvature or indicial). A given ametropia may be a combination of these two types. The size of the image on the retina of the ametropic eye corrected by an ophthalmic lens as compared to that of the normal eye, will be different in these two types of ametropia.

Now the total power ϕ of an ametropic eye of power ϕ_z combined with the correcting lens of power D, will be $\phi=D+\phi_z-D\phi_z s$, where s is the separation of the two systems, namely, the eye and the correcting lens (separation of the second principal plane of the former from the first principal plane of the latter). It so happens that

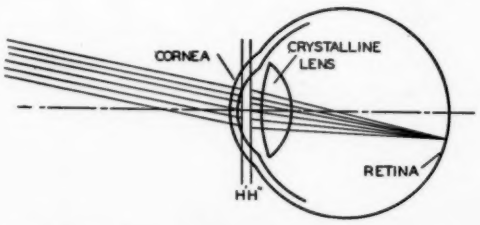


FIG. 1. Scheme of a normal or emmetropic eye.

* Invited paper presented at the symposium on Vision at the Optical Society winter meeting in New York City, March 5, 1943.

spectacles are worn at a position not far from the anterior focal point of the eye. If the separation s is made equal to the focal length of the ametropic eye, that is, $s=1/\phi_x$, ϕ will be equal to ϕ_x. Thus the power of the combined correcting lens and eye will be equal to the power of the ametropic eye itself. Usually variations in the position of the lens from the anterior focal point will introduce second-order errors.

The magnification of the retinal image of the corrected eye would then be $i_x/O=-1/(1-\phi_x u_x)$. The object distance u_x will be substantially equal to u.

Thus the ratio of the size of the retinal image in a corrected ametropic eye to that of the normal or emmetropic eye will be $R=i_x/i_0=(1-\phi_0 u)/(1-\phi_x u)$. If the object distance u is large, this ratio reduces to ϕ_0/ϕ_x, the ratio of the two powers.

Although the conditions under which the size of the retinal image in an eye corrected for a refractive error would or would not be different from that of the normal eye were stated as long ago as 1870,[1] they have not been fully understood.

Clearly, therefore, for the eye with a change in axial length the dioptric system remains unchanged and $\phi_x=\phi_0$ and $R=1$. This is the statement of Knapp's law, *viz.*, that if the second principal plane of a correcting lens is placed at the anterior focal point of the ametropic eye, the size of the image on the retina will be the same as though that eye were normal (emmetropic). (See Fig. 2.)

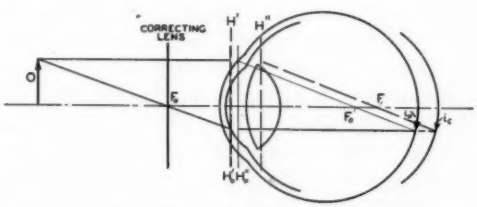


FIG. 2. Schematic axially ametropic (myopic) eye, showing that the retinal image is the same size as though the eye were emmetropic.

Just as clearly, in the case of the refractive ametropic eye, the size of the image of the corrected eye will be different from that image were the eye emmetropic. (See Fig. 3.)

This analysis can be carried one step farther by assuming that the "physiological size" of the retinal image, i.e., the size that is transmitted along the visual pathways, depends upon the linear number of retinal elements (which more broadly includes the entire visual processes) upon which the dioptric image falls. This number of elements, N, will be equal to the product of the linear density d of the "physiological retinal elements" and the size of the retinal image. Hence the ratio R above would become $R=N_x/N_0=(d_x\phi_0)/(d_0\phi_x)=(d_x/\phi_x)/(d_0\phi_0)$. Thus ultimately it will depend upon the ratio of the linear density of the retinal elements and the dioptric power of the eye, whether a given ametropic eye when corrected by an ophthalmic lens will act as an axial or refractive or a combined axial-refractive ametropia. Knapp's law as originally stated implies that $d_x=d_0$, that is, that the linear density of the retinal elements is the same in both the axial ametropic and emmetropic eye. There may be no reason for assuming otherwise.

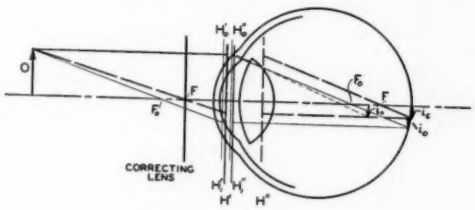


FIG. 3. Schematic refractive ametropic (myopic) eye showing that the size of the retinal image is different in the corrected eye from that if the eye were emmetropic.

One cannot know in any given case whether the ametropia is axial or refractive, except in certain cases of high myopia (near-sightedness), where observable structural changes in the retina, such as a characteristic crescent at the optic disk and straightening of the blood vessels, indicate a lengthening of the eyeball. Astigmatism obviously is a curvature ametropia in one meridian, the greater part of which can usually be measured directly by means of the ophthalmometer. On the other hand it is generally held that spherical ametropia is an anomaly in the length of the eyeball.

[1] H. Knapp, Arch. Ophth. and Otol. **1**, 377 (1870).

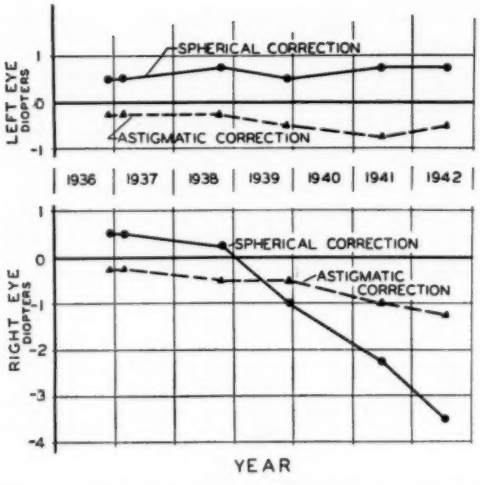


FIG. 4. Graph showing the increase in index myopia, in an illustrative case of refractive ametropia associated with a nuclear sclerosis of the crystalline lens of the right eye. (Reprinted by permission of the American Journal of Ophthalmology.)

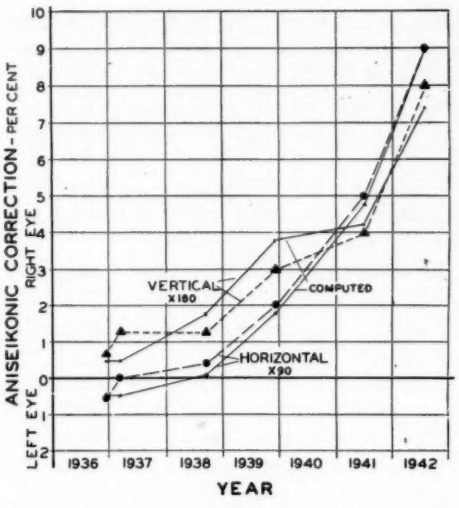


FIG. 5. Graph showing the increase in the measured and computed differences in the size of the images in the two eyes (aniseikonia) that accompanied the increase in index myopia of one eye in the illustrative case. (Reprinted by permission of the American Journal of Ophthalmology.)

Two clinical cases which illustrate the two foregoing types of ametropia are as follows:[2]

Not infrequently an increase in myopia is observed in certain patients over 60 years of age. This increase in near-sightedness is due to an increase in the index of refraction of the crystalline lens that accompanies a nuclear sclerosis of that lens (incipient cataract). A patient has been studied by our clinical group over a period of six years in which a myopia of this type developed in one eye only, the refractive error of the other eye remaining substantially unchanged during that period. The refractive errors of the two eyes during the six years are illustrated in Fig. 4, which shows the striking increase in near-sightedness of the right eye. The increasing myopia of the right eye is a refractive ametropia (index myopia), hence its correction would be expected to change the size of the retinal image. During the period of observation the difference between the sizes of the images in the two eyes (aniseikonia) was measured at each examination and an increasing image size difference proportionate to the increase in myopia was consistently measured. These measurements in the horizontal and vertical meridians are shown in Fig. 5. Because the increased myopic error took place in only one eye, and because we know that the origin of that error lies in the increased index of refraction of the crystalline lens, we can compute the change in the size of the image to be expected as the myopia increased. The calculated image-size difference to be expected is also shown on the graph. In these the calculations of the magnifying properties of the particular spectacles worn and the effect of the correction of the astigmatic error in each eye are taken into consideration. Agreement between the measured data and the calculated results is evident.

In a second case, in which there is a nearly identical difference in spherical refractive error between the two eyes to that of the first patient when last examined, no difference between the sizes of the images ($\pm\frac{1}{4}$ percent) was measurable. Moreover, the ophthalmoscopic examination of the retina of the myopic eye showed characteristic signs of a myopic stretching. Clearly this was a case of an axial myopia.

A survey of the literature and the textbooks on the subject of the origin of spherical refractive errors shows that these errors are generally considered to be axial. However, more recently,

[2] H. M. Burian and K. N. Ogle, Am. J. Ophthal. **26** (May, 1943).

several investigators[3] have given evidence showing that axial anomalies account for only a part of the spherical ametropias and that a combination of axial and refractive ametropias is even more frequent than simple axial ametropias for low errors. These results are based upon calculations made from measurements of the corneal and lenticular curvatures, etc., and from assumed indices of refraction of the ocular substances.

Figure 6 is a graph based on the data given by the Russian ophthalmologist, Tron, which shows the incidence of axial and refractive ametropias as computed by him. These calculations when compared statistically with similar computations from normal (emmetropic) eyes, show that combinations of axial and refractive ametropia account for a large proportion of the refractive anomalies, except for high myopia and high hyperopia. In general it is also pointed out that the variations in axial lengths found in cases of refractive anomalies are as great as the variations found in the axial lengths of emmetropic eyes. As an example, Fig. 7 is a graph showing the range of variations in calculated axial length of eyes for different refractive anomalies, based upon calculations by L. Paul, and using a somewhat different method than that used by Tron. The data shown in these graphs, while subject to the errors of measurement, calculation and interpretation, suggest that, except in cases of high near-sightedness or high far-sightedness, the refractive ametropia plays a prominent part in the origin of spherical errors.

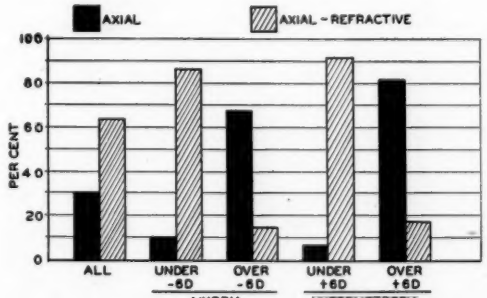


FIG. 6. Histogram illustrating data determined by E. Tron [Arch. f. Ophthal. **132**, 182 (1934)] on the incidence of axial and combination of axial and refractive ametropia in different degrees of refractive anomalies of the eye.

[3] E. Tron, Arch. f. Ophthal. **132**, 182–223 (1934); also Arch. f. Ophthal. **133**, 211–230 (1935). L. Paul, Arch. f. Ophthal. **133**, 254 (1935).

In this paper tentative data are presented for an empirical approach to the problem of the origin of the ametropias by means of a study of the measured difference in size between the ocular images of the two eyes in cases of spherical anisometropia; that is, in individuals where there is a difference in the spherical refractive error between the two eyes. Depending upon the nature and amount of the difference between the sizes in the images of the two eyes and upon the degree of spherical anisometropia, inferences may be drawn as to the origin of the anisometropia on the basis of the theory given above. The clinical staff of the Dartmouth Eye Institute has collected considerable data, which are available for analysis, on patients with anisometropia.

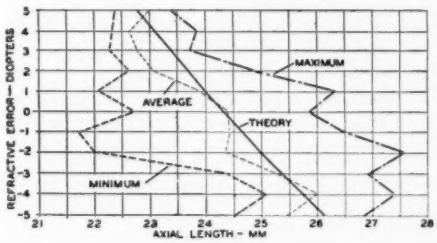


FIG. 7. Graph redrawn from that given by L. Paul [Arch. f. Ophthal. **133**, 254 (1935)] showing variation in computed axial length of eyes of different refractive anomalies.

For this study one hundred and thirty-one cases of spherical anisometropia were used. Unfortunately, in any such study the number of cases in which astigmatism is not present is very small, so that, in order to have a sufficient number of cases for statistical study, those with astigmatism must be included. In these cases it is necessary, however, first to eliminate from the measured image size correction that part which is due to the correction of the astigmatic errors. This can be readily done, since in all cases the data on the corneal astigmatism and the shapes and thicknesses of the spectacles and their distances from the corneas are available. From these data the relative magnifications of the dioptric systems comprised of the spectacle lenses themselves and of the spectacle lenses

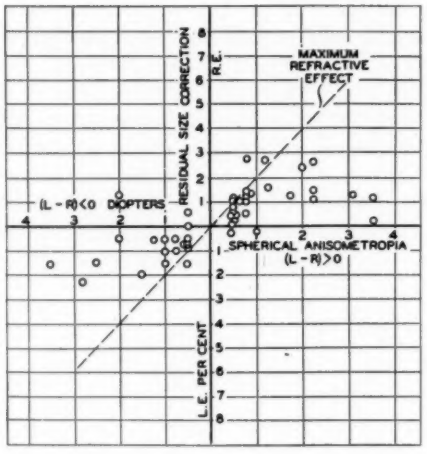


FIG. 8. Scatter diagram showing distribution of cases of spherical anisometropia and image size corrections (aniseikonia) when the astigmatic error in the two eyes differed by 0.5 diopter or less. Points along the abscissa indicate an axial anisometropia; points approaching the maximum effect line indicate a greater degree of refractive spherical anisometropia.

combined with the astigmatic error of the eyes can be computed.[4] Since the image size differences (aniseikonia) are measured in the horizontal and vertical meridians, the components of the computed magnifications due to the astigmatic correction in these meridians are likewise determined. From these the measured image size correction to be expected from the correction of the astigmatism can be found. In subtracting the computed values due to the astigmatic correction from the measured image size correction one obtains the unaccounted for or residual size difference correction. (Cf. Table I.)

If these residuals in the horizontal and vertical meridians did not differ by more than $\frac{1}{4}$ percent or $\frac{1}{2}$ percent image size difference from their mean value (depending upon the precision of measurements), the residual image size difference correction was considered an over-all difference and assumed to be probably correlated with the spherical anisometropia present. If the residual corrections in these meridians in a given case did not agree within these limits, they were considered to be due to other than simple dioptric factors and such cases, not being directly associated with spherical anisometropia, were not used in this study. The cases in this latter group comprised about 35 percent of the cases studied.

In the determination of the amount of spherical anisometropia, a fundamental difficulty exists which cannot be easily avoided in cases where an appreciable astigmatic difference between the two eyes is present. We have no simple way of determining from the prescription as written, whether the astigmatism is due to a flattening of some ocular surface in one meridian (+ cylinders) or an increased curvature in the meridian at right angles (− cylinders), each combined with a different spherical ametropia. As an example, the refractive correction of a given subject (case 205 in Table I) can be written in either of two ways (really four but the usual symmetry of corneal curvatures in anisometropia suggests two as rather improbable).

(1) R -2.00 sphere
L $+1.25-0.75\times45$,

(2) R -2.00 sphere
$+0.50+0.75\times135$.

In the first case the spherical anisometropia (L.E.−R.E.) is +3.25 diopters; in the second

TABLE I. Table showing typical cases analyzed in the study of the relationship between spherical anisometropia and residual correction for the measured difference in size between the images of the two eyes.

No.	Refraction	Spherical anisometropia (L-R)	Analysis	Percent image size correction Horiz.	Vert.
2	R $+5.00-0.50\times180$	-2.50	Meas. L	2.5	L 2.5
	L $+2.50-1.00\times3$	(-3.00)*	Astig. L	0.7	L 1.4
			Resid.	−1.8	−1.1
45	R $+0.75-2.50\times180$	$+0.50$	Meas. R	0.1	R 0.3
	L $+1.25-2.50\times12$	$(+0.50)$	Astig. R	0.1	R 0.2
			Resid.	0.0	+0.1
71	R $+1.75$	$+2.25$	Meas. R	2.0	R 2.0
	L $+4.00-0.50\times65$	$(+1.75)$	Astig. L	0.7	L 0.5
			Resid.	+2.7	+2.5
92	R -0.75	-1.00	Meas. L	2.0	L 2.0
	L -1.75		Astig.	0.0	0.0
			Resid.	−2.0	−2.0
127	R $-3.50-0.75\times5$	$+2.25$	Meas. R	1.5	R 1.5
	L $-1.25-0.75\times165$	$(+2.25)$	Astig. L	0.1	R 0.1
			Resid.	+1.6	+1.4
205	R -2.00	$+3.25$	Meas. R	2.2	R 2.6
	L $+1.25-0.75\times45$	$(+1.50)$	Astig.	0.0	L 0.2
			Resid.	+2.2	+2.8
251	R $-0.25-1.25\times170$	-2.00	Meas. L	1.3	0.0
	L $-2.25-1.25\times45$	(-2.00)	Astig. L	0.8	R 0.5
			Resid.	−0.5	−0.5

* ( ) the spherical anisometropia when the plus cylinder notation is used.

[4] K. N. Ogle, J. Opt. Soc. Am. **26**, 323 (1936).

case +2.50 diopters. Thus in a number of the cases available for study the exact amount of spherical anisometropia is equivocal.

The easiest way to circumvent this difficulty was to select cases in which the astigmatic errors in the two eyes differed as little as possible, specifically not more than 0.5 diopter. In this group the spherical anisometropia, which will be expressed as the difference of the spherical refractive errors of the two eyes (L.E.−R.E.), would be substantially the same regardless of the manner in which the astigmatic errors are considered. The data falling in this category were plotted in a scatter diagram, Fig. 8, where the abscissa represents the spherical anisometropia in diopters and the ordinate, the residual or unaccounted for over-all image size difference correction in percent.

The maximum magnification or reduction in the size of the retinal image due to the correction of a refractive ametropia would occur when the origin of that ametropia is as far back in the eye as possible, *viz.*, in the crystalline lens. This expected maximum residual image size correction can be computed, assuming thin-lens spectacles placed at an average distance from the eyes. The dashed line in the graph indicates that maximum. The points representing the data for all the anisometropic cases in which the spherical ametropic difference was due to an axial anomaly would be expected to lie near the abscissa, within the precision of the measurement. Similarly, if the spherical ametropic difference is refractive in nature, the maximal data would lie along the dashed line. Points lying between would indicate combinations of axial and refractive types.

Thus *a priori*, one would anticipate that all the points representing the data would lie in the regions of the first and third quadrants of the graph bounded by the abscissa and the dashed line. It is not surprising to note that the majority of points representing the data tend to show just that.[5]

The data from the remaining cases of anisometropia can be included if the value of the spherical

[5] In view of the fact that a certain percentage of the anisometropic cases studied had residual size corrections which could not be attributed wholly to spherical anisometropia, it is probable that there would be some cases in which even the residual size correction though over-all might be due to factors other than dioptric.

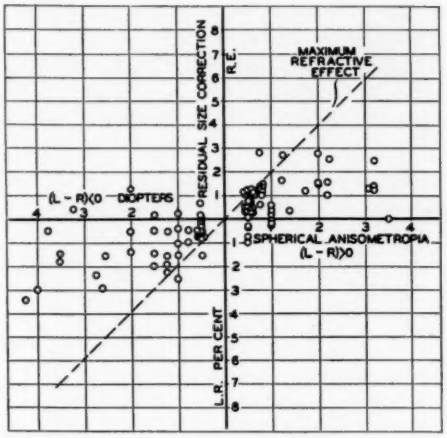


FIG. 9. Scatter diagram of available cases showing distribution of cases of spherical anisometropia and the residual image size difference (aniseikonia) corrections when the effect of the correction of the astigmatic errors is eliminated. Points along the abscissa indicate an axial anisometropia; these approaching the maximum effect line indicate greater degrees of refractive spherical ametropia.

anisometropia can be determined in each case. In some this can be decided through a comparison of the corneal curvatures with the astigmatism in the two eyes, by assuming that the corneal meridians of equal curvature represent the basic condition. If one plots in the scatter diagram the points for each of the two equivocal spherical anisometropias, in a number of cases one of the points will fall outside the critical region of the graph, thus suggesting that the other designation was the correct one. In many of the cases, however, both points fall within the critical region of the graph so that the value of spherical anisometropia must indeed remain equivocal. However, using either of the designations does not change the essential character of the diagram.

Figure 9 is the diagram for all the data available, the equivocal cases being plotted in terms of the astigmatism as expressed in minus cylinders (suggested by the much higher percentage of cases where the spherical anisometropia was not equivocal). Inspection of this figure essentially strengthens the conclusions drawn from Fig. 8. Very roughly one can estimate from these scatter diagrams that in 30 percent of the cases the spherical anisometropia was axial in nature, in 30 percent was essentially refractive in nature,

and in 40 percent was a combination of axial and refractive ametropias.

These results indicate, therefore, that refractive ametropia, at least in part, is a cause of the spherical ametropic errors where the spherical anisometropic difference is less than four diopters. Except that a higher proportion of refractive ametropia is found here, this conclusion would be supported by the results of Tron.

These results are presented at this time only tentatively, the primary purpose being to suggest a more empirical method of approach to the problem. In the first place the results are restricted by the relatively small number of cases available and in the second place, there may be a question whether the population group of anisometropic cases studied so far is a representative sample. Irrespective of these limitations, however, the method is of considerable interest and may prove fruitful.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 33, NUMBER 9 SEPTEMBER, 1943

A Single Judgment Test for Red-Green Discrimination[1]

LEGRAND H. HARDY[2]
Institute of Ophthalmology, New York, New York

MATERIAL developed for other ISCC committee work[3] provided two series of small-step chroma differences in hues designed to confuse the observer who makes errors of the typical protanope and of the typical deuteranope in accordance with available data, chiefly those published by Pitt.[4] The series extend from Munsell N 5/ in 40 steps of saturation to 6 R 5.4/4 for the protanope series, and in 40 steps of saturation to RP 5.0/4 for the deuteranope series. The material is in the form of metal chips, $1\frac{1}{4}\times1\frac{1}{2}''$, the color baked on an enamel surface, the pigments used being as permanent in color as any that are known.

Several forms of simple screening tests using this material have been developed, but in early stages of the work representatives of the Armed Forces insisted that a satisfactory test could have no loose parts. The material has been tied down therefore to a fixed mechanism (Figs. 1 and 2). In this device 20 pairs of chips are mounted juxtaposed on a disk 15″ in diameter. The mask has two openings, one at the left, one at the right, each of which exposes one pair of chips at a time. The subject is asked to indicate whether the left chip of a pair or the right chip is the redder. In the model illustrated the subject needs only to press a left button or a right button, thus avoiding the possibility of confusing right and left, and in the Navy model a shutter over the chips automatically exposes each pair for a given number of seconds. The effect of having two openings in the mask is that 40 judgments are obtained from 20 pairs of chips, and that the right-left position of the members of each pair are taken into account.

The color of illumination is highly critical for color vision tests, therefore standard artificial daylight (Macbeth) in the color temperature range of 6700°K–7500°K has been made a part of this test device. Illumination as low as 10 ft.-c seems to be satisfactory, but since this requires that the test be administered in some sort of a dark room, it is more feasible, in order to maintain test conditions in a room only partially darkened, to use a higher illumination. Much of the present test work is being done, therefore, with illumination of 20–60 ft.-c. Illumination differences in the range between 10 ft.-c and 60 ft.-c do not seem to be critical.

During the first several months of work on this test (February, 1942–October, 1942) our co-

[1] This paper is part of a symposium on Color Blindness and Color Blindness Tests arranged for the March 4–6, 1943 meeting of the Optical Society of America by the Inter-Society Color Council.

[2] Chairman, ISCC Committee on Development of a Red-Green Discrimination Test.

[3] Forrest Lee Dimmick, J. Opt. Soc. Am. **32**, 745(A) (1942).

[4] F. H. G. Pitt, Med. Res. Council (Brit.) Special Report Series, No. 200 (1935); F. H. G. Pitt, Proc. Phys. Soc. **14**, 219–244 (1942).

operating members used somewhat divergent forms of the test and test procedure. In October, after studying the results, it was decided to adopt standardized test procedure. Since that time the data collected by most of the committee members have been in accordance with standard instructions. The order of size of color difference, of right-left, of protanope and deuteranope chip pairs, is now random. A standard "warming-up" procedure has been adopted: The subject reads a card of instructions on which sample pairs are shown. Thus all subjects get identical instructions; there should be no necessity for verbal instruction.

The Pitt chromaticity confusion lines of the typical protanope and deuteranope, transferred to the I.C.I. (x, y) diagram, are illustrated in Figs. 4a and 4b of an article by Judd in this journal.[5] A portion of this diagram with the protanope and deuteranope lines is illustrated in Fig. 3, the heavy line representing colorimetric data for the materials used in this test.

Tests made preliminary to standardization included a series on 57 near-normals (seamen of the British Navy) tested at the Ophthalmological Institute of the Presbyterian Hospital, 29 color-deficients tested at Hobart College, 396 men tested by the Navy, a general population of recruits for Air Corps cadets plus a few known color-deficients, 10 color-deficients tested at Randolph Field, and 6 color-deficients tested at the United States Department of Agriculture. Since standardization, we have assembled tests made on color-deficients at the Ophthalmological Institute, at Hobart College, and by Mr. Farnsworth of New York University. Following the next committee meeting we expect to assemble returns on tests made since adoption of the standard test method.

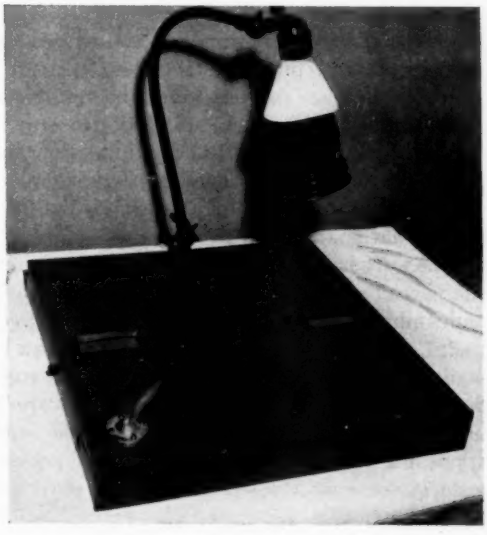


FIG. 1. One model of ISCC red-green discrimination test, ready for testing.

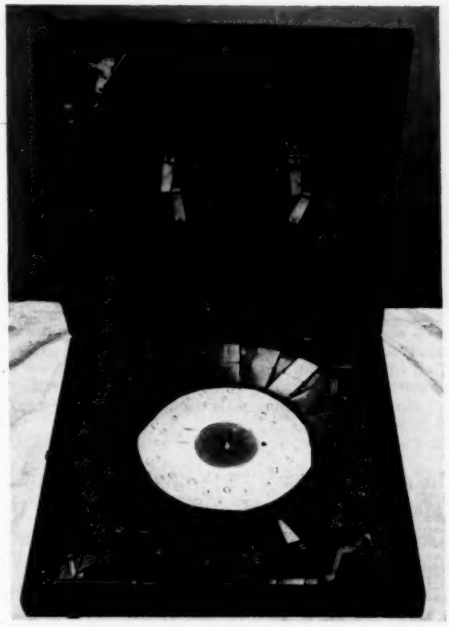


FIG. 2. Inside view of model in Fig. 1 showing paired chips in place on moveable disk.

Scoring. Since the yellowish-red series of chips is devised to confuse the protanope, the typical protanope has an even chance of calling each of the yellowish-red chips right or wrong. The deuteranope has an even chance of calling the purplish-red series right or wrong. Just as one expects, the majority of red-green blinds will make 9, 10, or 11 errors out of 20 presentations. Scores are recorded as protanope and deuteranope error scores. If a subject makes less than 10 percent error in either series, he is judged normal.

[5] Deane B. Judd, J. Opt. Soc. Am. **33**, 305 (1943).

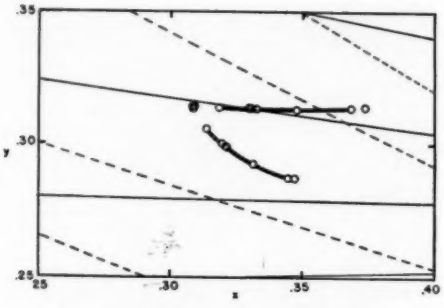


FIG. 3. Portion of I.C.I. (x, y) diagram. Continuous black lines present Pitt chromaticity confusion lines of protanope, dotted lines those of deuteranope. Circles represent measurements of selected samples of ISCC material, with heavy black line indicating materials used in the test described in this paper.

If he makes errors of 35 through 65 percent, he is judged R-G blind. If errors lie between 10–35 percent in either or both series, he is judged deficient in R-G discrimination. If errors are greater than 65 percent, it is probable that the subject misunderstood instructions, confused right and left, confused redness with lightness-or-darkness for the protanope series, or is a malingerer. When the scores are definitely near to chance for one series only, then the subject is indicated as a protanope or deuteranope if near to a 50 percent score—protanomalous or deuteranomalous if near to a 25 percent score. If scores in both series run about the same then the subject is called "R-G blind, undetermined type" if near to 50 percent, the "R-G deficient, undetermined type" if nearer 25 percent.

Repeated tests on the same observer, tests with the same observer on different sets of chips, on the same observer with different methods of test, have been made and results assembled by the committee.

Each subject is tested not only on the ISCC test but on the A.O., the fifth edition of Ishihara, and the Farnsworth-Munsell 100-hue test, as well as other tests that the cooperating members have available. Some have included Rabkin, the new Meyrowitz, Jennings. And on the Navy series, test results are available for the ISCC, A.O., and the test described by Dr. Hecht.

There is only very general agreement between various test results, each test showing peculiarities in a few cases. For example, in a comparison of about 60 deficient subjects tested on the A.O. and the ISCC, 4 high A.O. error scores show up in low ISCC error scores, but on the other hand, 7 low A.O. scores have high ISCC error scores, some of them very high, which would indicate that a number of R-G anomalous and R-G blind individuals are being passed by the A.O. test. There are peculiarities in all tests, and we need to find out the cause for each of them before the job is done.

We believe we have gone far enough so that new test material and several copies of the device should be thoroughly tried out on general populations and on populations in which the Armed Forces are interested, particularly the Air Forces with whose cooperation the work has so far been done.

The test is made of durable and permanent materials to exact colorimetric specifications; it is simple and rugged, fairly convenient in size, portable; it can be made self-administering; it includes its own standard daylight; on the basis of 40 judgments, subjects can be divided into three groups; the test cannot be memorized. All of these things, plus the fact that in a large majority of cases the test gives the results expected of it, make it seem a useful one to develop further. So far only the R-G discrimination portion of the test has been developed. Original plans call for the inclusion of Y-B and lightness discrimination.

This work has been the result of cooperation on the part of many ISCC delegates and members, and while we feel that this test can be made very useful, we are just as sure that other tests can be made useful also. Once we know just what color discriminations are required by the various branches of the Armed Services, then the most applicable tests can be selected out of present tests, or new ones developed for specific purposes.

An Improved Electrode Holder for Spectrographic Analysis*

BOURDON F. SCRIBNER AND CHARLES H. CORLISS
National Bureau of Standards, Washington, D. C.

MANY of the parts of the spectrographic assembly have undergone marked improvement during recent years, but relatively little attention has been given to the design of the electrode holder, except for certain special applications. In addition to the common models based on the simple Gramont design, holders have been devised for a.c. arc excitation of rods[1,2] or sheet metal,[3] and for rapid change of spark electrode rods[4,5] or metal plates.[6] A universal holder has been described[7] that provides a variety of electrode movements and adjustments.

For application in miscellaneous analyses, however, difficulty may be encountered when changes must be made quickly to handle electrodes of different sizes and shapes and from one extreme of excitation conditions to another. To meet these demands, a holder has been constructed incorporating the following features: (1) ease and speed of operations over the varied conditions encountered, (2) precise adjustment and motion of parts to prevent errors caused by poor alinement of electrodes, (3) water-cooling of electrode clamps to remove excessive heat, thus preventing deterioration of the parts and, at the same time, cooling the electrodes where necessary, (4) provision for adjustment of distance of holder from the spectrograph slit by an optical bench to permit variation in illuminating conditions, and (5) a hood for protection against light, fumes, and electric shock.

The design of the holder to be described is based on the suggestions of several operators and incorporates improvements found advisable by tests on a preliminary model.

DESIGN AND CONSTRUCTION

A view of the holder, with the housing opened to expose the parts, is given in Fig. 1. Details of its design and construction are shown in the assembly drawing reproduced in Fig. 2. The holder consists of a heavy steel base *1*, standing on three legs (two *V*'s and a flat) that rest on the optical bench. The base is locked to the bench by a claw *2*, and is fitted with a slide that is moved transversely by control *3*. The slide carries two steel upright rods *4* that guide the electrode supports *5*. Each support consists of an *H*-shaped brass yoke, constructed from tubing and block brass, with a forward extension upon which is bolted a Bakelite plate *6*, carrying the electrode clamp *7*. The uprights *4*, are braced at the upper ends by a plate *8*, on which is mounted the upper

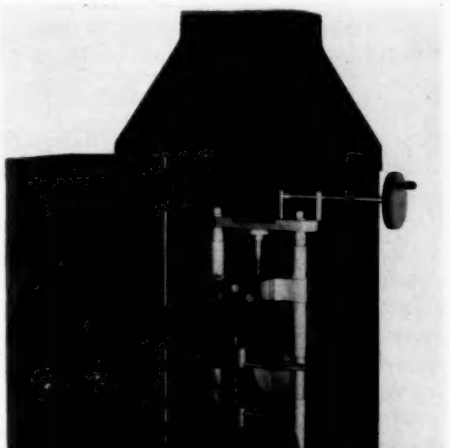


FIG. 1. Improved electrode holder.

* Presented at the Tenth Annual Conference on Spectroscopy and Its Applications, Massachusetts Institute of Technology, July 20–22, 1942.

[1] O. S. Duffendack and R. A. Wolfe, Ind. Eng. Chem., Anal. Ed. **10**, 161 (1938).

[2] W. D. Owsley and R. C. McReynolds, Rev. Sci. Inst. **13**, 342 (1942).

[3] R. A. Sawyer and H. B. Vincent, J. App. Phys. **11**, 452 (1940).

[4] Carl Zeiss, German Patent 681,294 (Aug. 31, 1939).

[5] H. B. Vincent and R. A. Sawyer, *Proceedings of the Seventh Summer Conference on Spectroscopy and Its Applications* (John Wiley & Sons, New York, 1940), p. 16.

[6] H. V. Churchill and J. R. Churchill, J. Opt. Soc. Am. **31**, 611 (1941).

[7] A. Gatterer, Ric. Spettroscopiche **1**, 55 (1939).

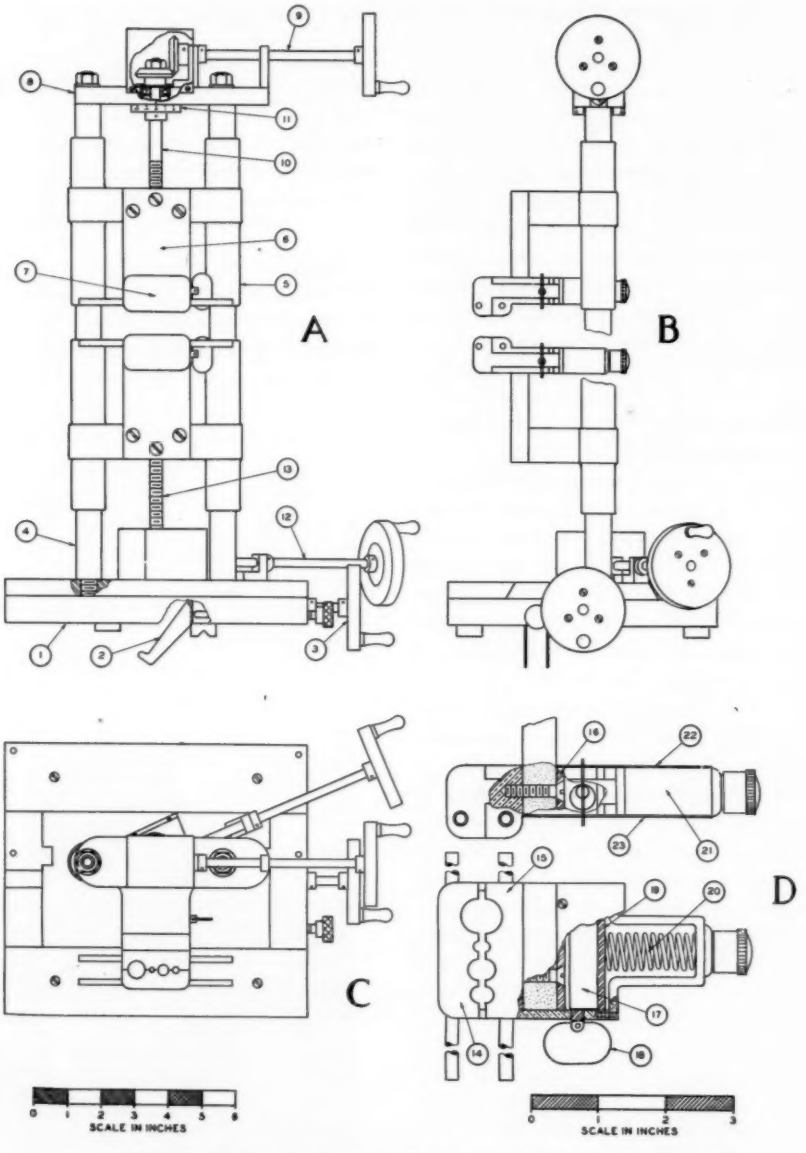


FIG. 2. Assembly drawing of the electrode holder. *A*, front view; *B*, side view; *C*, top view; *D*, detail of electrode clamp.

control assembly *9*. This control provides vertical movement of the upper electrode support through miter gears and a screw *10* threaded into the support. The screw chosen for this holder has a pitch of 1.5 mm. Precise setting of small electrode distances is provided by a drum *11*, mounted on the screw and graduated with 15 divisions, each division corresponding to a 0.1-mm travel of the electrode support. Independent motion of the lower electrode support is provided by a control *12*, with gears and a screw *13*, similar to the upper electrode assembly. The exposed metal parts,

except the electrode clamps, are nickel-plated to resist corrosion.

Enlarged views of one of the electrode clamps are shown in section *D* of Fig. 2. The clamp jaws are made of two pieces of copper, one forming the movable jaw *14*, and the other the fixed jaw *15*. The motion of the jaw *14* is precisely limited and controlled by a slide and ways on each side of the clamp. A quarter-turn of the knob *18*, provides the necessary opening and closing of the jaw ($\frac{1}{16}$ in.) through an eccentric shaft *17*. A spring *20* provides a fixed tension on the jaw for clamping the electrodes. The clamp assembly extends through the opening between the holder uprights and terminates in a binding post for connection to electrical power. The back cover *22* of the eccentric and spring box is Bakelite, and the face cover *23* exposed to the source is made of brass.

The electrode jaws are drilled to accommodate a varety of electrode diameters for the lower and upper clamps, respectively, as follows: $\frac{1}{2}$ in. *vs.* $\frac{1}{2}$ in., $\frac{1}{4}$ in. *vs.* $\frac{1}{8}$ in., $\frac{1}{4}$ in. *vs.* $\frac{1}{4}$ in., and $\frac{7}{32}$ in. *vs.* $\frac{7}{32}$ in. Slight variations from the standard electrode sizes cause no difficulty in clamping if the holes are made slightly oversize. Split copper bushings, machined to fit the jaws, permit other electrode combinations. The jaws of the clamp, once accurately alined, have maintained alinement during continuous use. All corners and edges of the jaws are rounded and the parts streamlined to facilitate cleaning. Copper tubes inserted through the jaws serve for water circulation. Finally, an accessory clamp which accommodates flat bar, disk, or sheet samples, may be quickly locked into either the upper or lower electrode clamp.

The electrode jaws are cooled by water circulated from a reservoir (5-gal. bottle) by means of a small rotary pump, the water flowing successively through the four jaws of the clamps before returning to the reservoir. Connections in the water system are made by gum-rubber tubing and glass. This, combined with the use of distilled water in the system, is effective in preventing leakage of high voltage (35,000-v peak) power. The closed water system, always at room temperature or slightly higher, ensures freedom from condensation of moisture on the holder when not in use. As a convenience in operation, the pump motor is connected to power through a master switch which closes electrical circuits to the spectrograph controls, a rotating sector, and a ventilator. Under these conditions the water-cooling system requires little attention from the operator.

The housing of the holder is built with a frame of $\frac{1}{2}$-in. angle iron and a covering of galvanized sheet iron, and is finished in black crackle enamel. The forward half of the housing, which is supported on hinges at the side, forms a door to provide access to the holder. When the door is opened, a switch and safety circuit shut off the power circuit to the electrode clamps and a second switch closes a circuit to a lamp mounted in the corner of the door. The door opens to a stop so arranged that the lamp is placed in a position to project an image of the electrodes to a screen on the optical path for adjustment of the electrode positions. The lamp serves also to illuminate the interior of the housing.

Openings are cut in the front and back of the housing for the passage of light along the optical path. Ventilation is provided by a hood at the top of the housing, connected through a flexible tube to a ventilating system. A small upward current of air is sufficient; excessive air currents will make an arc discharge unsteady or even extinguish it. The housing is grounded through the base of the holder and the optical bench.

APPLICATIONS OF THE HOLDER

The holder has been employed for spark excitation with the various electrode combinations employed in the Bureau's spectroscopic laboratory. These include the spark between metal rods of sizes from $\frac{1}{8}$ in. to $\frac{1}{2}$ in. in diameter, between $\frac{1}{8}$-in. graphite rod and flat metal specimens, and between pressed pellets $\frac{1}{4}$ in. in diameter and $\frac{1}{4}$ in. to $\frac{1}{2}$ in. long. Cooling of the spark electrodes is useful when small electrodes or sheet-metal specimens are sparked. Sheet-metal samples may be mounted either vertically in the slot between the jaws of the lower clamp or held horizontally in the accessory holder, with a supporting block of copper for rapid heat conduction. The sample in this case is opposed by a counter electrode of graphite. The holder has been found advantageous in establishing a reproducible condition of heat conduction from the electrodes. With holders previously used in this laboratory, con-

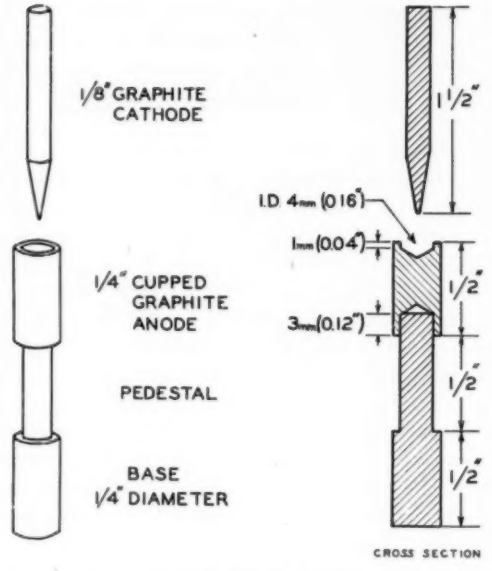


FIG. 3. Electrode assembly for direct-current arc excitation.

siderable variation in conduction has been encountered, especially after deterioration of the holders in use.

Excitation by the a.c. arc generally requires close control of alinement and gap distance of electrodes. In order to set the electrodes accurately for small gaps, the upper electrode is lowered almost in contact with the lower electrode, the clamp is opened slightly, permitting the upper electrode to drop into contact with the lower, the clamp is then released to grip the electrode, and, finally, the electrode is raised the required distance, which is measured on the graduated drum (Fig. 2, part *11*). The holder was found satisfactory not only in meeting the requirement of accuracy in the settings, but also in providing the advantage of cooling the electrodes. Small pellets, or $\frac{1}{8}$-in. rod-samples may be arced with little danger of over-heating.

Direct-current arc excitation results both in radiation of heat and in evolution of hot, corrosive vapors. The design of the electrode clamps and efficient water cooling have proved effective in meeting these conditions for rapid and continuous use of the holder. The cooling of electrodes of various metals in the form of compressed pellets permits use of currents as high as 10 amp. without danger of melting the electrode except at the end. However, if heat is required in the electrode, as in volatilizing samples from a hot graphite anode, unsatisfactory results may be observed with the customary electrode shapes. As a result of the high cooling efficiency of the holder, the electrodes are attacked only on the extreme ends with currents as high as 15 amp. To overcome this difficulty, the d.c. arc electrode assembly was modified to provide an electrode combination that has proved adaptable to a number of applications of the d.c. arc, both for the water-cooled holder and common clip holders. The assembly, with dimensions shown in Fig. 3, consists of a $\frac{1}{8}$-in. upper cathode tapered to a point and a lower anode combination in two parts—first, a graphite base and pedestal supported in the electrode clamp; and, second, the electrode cap. The cap is drilled at both ends, the depth of the upper cup depending on the degree of heating required and the size of sample charge. By virtue of the low heat conductivity between the cap and the pedestal, the cap becomes very hot with only slight attack on the pedestal. This electrode combination gives reproducible excitation and requires a smaller quantity of graphite rod, which is a welcome economy when expensive high-purity electrodes must be used.

Experience with two of the improved electrode holders for numerous tests under various conditions shows that they have increased the speed of miscellaneous testing, provided more reproducible electrode orientation, and contributed to the safety and ease of operations.

The authors acknowledge the helpful suggestions of H. R. Mullin, of the Spectrographic Laboratory, and of Henry Philo of the Bureau's Instrument Shop.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 33, NUMBER 9 SEPTEMBER, 1943

Computation of Polynomial Functions by Summation of Finite Differences

H. F. BENNETT
Hawk-Eye Works, Eastman Kodak Company, Rochester, New York
(Received May 14, 1943)

A quick approximate method is described for use with a calculating machine in computing a table of values of a polynomial function which itself may be an approximation to another function. Formulas are given for polynomials up to the eighth degree whereby the computation may be based either upon the coefficients of the polynomial or upon a few equally spaced values of the function. Formulas are given also for predicting the maximum deviation from the exact values of the polynomial. The method is essentially a computing routine which makes use of an arbitrarily defined array of differences and results in replacing the given curve by a series of overlapping parabolic arcs, each extending through five consecutive values of the argument.

1. INTRODUCTION

THE computing method herein described was developed for computing aspheric lens curves, but it is equally useful in computing cams and similar curves and perhaps in computing some types of mathematical tables. Although the computed numerical values are not theoretically exact, a sufficiently close approximation is obtained for many ordinary purposes and with a great saving of time.

The author feels that the most important part of the paper is the computing routine (§ 4, 5, and 6), which forms a Procrustean bed into which a simple elegant theoretical treatment of the method is very difficult to fit. A distinction must be drawn between the practical and the theoretical aspects in that certain difference values (e.g., $Z_{II,1}$ and $Z_{II,2}$) which exist theoretically are omitted from the written work but are implicit in the computing routine.

From the theoretical standpoint, the periodic deviation from the original polynomial is such that the second difference remains constant for three successive values. This second difference is itself then regarded as a function to be computed for values one-third as frequent as the (approximate) values of the original polynomial, and a further deviation from the original polynomial is caused by imposing the added condition that the second differences of these second differences of the approximate function must remain constant for each set of three successive values. In order to avoid the ensuing complications of nomenclature, the author has perhaps done violence to the conventional meaning of the word "difference" and simply refers to the first, second, third, and so on, orders of "modified differences," or, if not ambiguous, simply "differences."

From the practical standpoint, only one out of every three of the theoretical values of the modified differences is written down. In other words, one nth order difference is written for every $3^{(n+1)/2}$ or $3^{n/2}$ values of the argument, whichever is an integer.

2. DEFINITIONS

If Y is a polynomial function of X, e.g.,

$$Y=3+X+2X^2$$

and values of Y are tabulated for equally spaced values of X, the first differences are then customarily defined as the numbers found by subtracting each Y from the next, and second differences by subtracting succeeding first differences, and so on, as shown in Table I.

If there has been no rounding off of decimals, the nth difference is constant, where X^n is the

TABLE I. Classical array of differences for $Y=2X^2+X+3$.

X	Y	First diff. Δ_I	Second diff. Δ_{II}
0	3		
		3	
1	6		4
		7	
2	13		4
		11	
3	24		4
		15	
4	39		4
		19	
5	58		

highest power of X in the function. If some decimals have been dropped, however, some order of differences will be reached in which the successive values are irregularly plus and minus. A more complete treatment of the subject may be found in nearly any textbook on statistics[1] or numerical analysis.[2]

3. COMPUTING THE FUNCTION FROM THE DIFFERENCE

Obviously, further first differences can be computed in Table I by repeated addition of the second difference (a constant in this case), and likewise the function itself can be found by adding the successive first differences thus found. This process may, of course, be extended to polynomials of higher degree, the computer often starting with only the top number (known as the leading term or leading difference) in each column. It is a modification of this classical[3] method that is proposed in the present paper.

4. NEW METHOD APPLIED TO THE ORDINARY PARABOLA

According to the modified method proposed here, time is saved by omitting two out of every three of the first and second differences from the written work. Of the possible ways of doing this, the pattern shown in Table II was chosen as embodying the most symmetrical arrangement.

TABLE II. New method applied to the parabola, $Y=2X^2+X+3$.

X	Y	$Z_I=\Delta_I$	$Z_{II}=\Delta_{II}$
0	$3=Y_0$		4
1	$6=3-4+7$	7	
2	$13=6+7$		
3	$24=13+7+4$		4
4	$39=24-4+19$	$19=7+3\cdot4$	
5	58		
6	81		4
7	108	31	
8	139		

As a simple example, the same parabola shown in Table I is shown computed by the new method in Table II. Since the symbol Δ has a very definite conventional meaning, a different symbol, Z, is adopted here. It will develop later that Z and Δ are not the same except in the case of first and second-degree equations. The roman numeral subscript indicates the order of differences, and the arabic subscript designates which difference of any given order is meant. This arabic subscript is arbitrarily chosen as the value of X on the same line as the difference designated.

The routine on the calculating machine is as follows: As many successive values of Z_I as desired are computed by repeatedly adding $3Z_{II}$. Then starting with $Y_0=3$ in the register of the machine, subtract $Z_{II}=4$, then add $Z_I=7$, leaving the 7 in the keyboard. These two operations are equivalent to adding $\Delta_{I,\frac{1}{2}}=3$ as in the classical method. Again add 7 and again leave it in the keyboard. This gives $Y=13$ corresponding to $X=2$. Next, add the 7 a third time, clearing the keyboard, and add $Z_{II,\,3}=4$, leaving 4 in the keyboard. The latter two operations are equivalent to adding $\Delta_{I,\,2\frac{1}{2}}=11$ in the classical method. The next cycle of three Y values may be obtained in the same way, using $\Delta_{I,\,4\frac{1}{2}}=19$, but with this advantage, that the $Z_{II,\,3}=4$ is already in the keyboard.

This method gives exact results for a second-degree equation except for rounding off of decimals, and its advantages over the classical method become quickly evident when there are several significant figures in Δ_I and Δ_{II}.

TABLE III. Leading differences expressed in coefficients of polynomial.

$$
\begin{aligned}
Z_{VIII,\,0} &= 2\ 14277\ 01120J \\
Z_{VII,\,4\frac{1}{2}} &= 1\tfrac{1}{2}Z_{VIII,\,0}+992\ 02320H \\
Z_{VI,\,0} &= 5\ 24880G-6686\ 97120J \\
Z_{V,\,1\frac{1}{2}} &= 1\tfrac{1}{2}Z_{VI,\,0}+9720F+155\ 13120H \\
Z_{IV,\,0} &= 216E-16200G+7954\ 37496J \\
Z_{III,\,4\frac{1}{2}} &= 1\tfrac{1}{2}Z_{IV,\,0}+18D+1530F-113\ 39622H \\
Z_{II,\,0} &= 2C-6E+19242G-31763\ 68326J \\
Z_{I,\,1\frac{1}{2}} &= 1\tfrac{1}{2}Z_{II,\,0}+B+3D-1149F+895\ 86603H \\
Y_0 &= A
\end{aligned}
$$

[1] For example, Rietz and others, *Handbook of Mathematical Statistics* (Houghton Mifflin, Boston, 1924), 34ff.

[2] For example, Scarborough, *Numerical Mathematical Analysis* (The Johns Hopkins Press, Baltimore, 1930).

[3] The author has been unable to determine the origin of this method. It seems to have been commonplace among 18th century mathematicians. Babbage began building a difference engine in or about 1822 for the purpose of constructing tables by this method, but his machine was never completed.

TABLE IV.

X	$Z_V \times 10^6$	$Z_{IV} \times 10^4$	$Z_{III} \times 10^4$	$Z_{II} \times 10^2$	$Z_I \times 10^2$	$Y = U$	V	Error $\times 10^5$
0		−0.060750		+0.012516		+5.68750	5.00	
1					+9.70460	+5.78442	5.05	
2						+5.88147	5.10	
3				+0.002079		+5.97853	5.15	
4			−1.10448		+9.71084	+6.07562	5.20	
5						+6.17273	5.25	
6				−0.008966		+6.26975	5.30	
7					+9.68394	+6.36668	5.35	
8						+6.46352	5.40	
9		−0.066217		−0.020673		+6.56015	5.45	
10					+9.62192	+6.65658	5.50	
1						+6.75279	5.55	
2				−0.033042		+6.84868	5.60	
3						+6.94424	5.65	
4	−0.54675		−1.30313		+9.52279	+7.03947	5.70	
15				−0.046074		+7.13424	5.75	
6					+9.38457	+7.22854	5.80	
7						+7.32239	5.85	
8		−0.071685		−0.509822		+7.41564	5.90	
9					+9.20511	+7.50829	5.95	
20						+7.60034	6.00	
85						+9.20121	9.25	
6			−3.48285		−10.10712	+9.10014	9.30	+3.
7				−0.610623		+8.99296	9.35	
8						+8.87968	9.40	−4.8
9					−11.93899	+8.76029	9.45	+3.2
90		−0.115425		−0.646606		+8.63443	9.50	−1.
1						+8.50211	9.55	−5.2
2					−13.87881	+8.36332	9.60	−1.
3				−0.683743		+8.21769	9.65	−1.8
4						+8.06523	9.70	−5.8
95	−0.54675		−3.82912		−15.93004	+7.90593	9.75	+2.7
6				−0.722034		+7.73941	9.80	−1.6
7						+7.56567	9.85	−5.8
8					−18.09614	+7.38471	9.90	+3.
9		−0.120892		−0.761534		+7.19613	9.95	−2.
100						+6.99994	10.00	−6.
1					−20.38074	+6.79613	10.05	

5. THE NEW METHOD ADAPTED TO EQUATIONS OF HIGHER DEGREE

In computing polynomials of degree higher than the second, a deviation is made from the exact theoretical values. The deviation is such that the second difference remains constant for three successive values of the argument. However, it is entered only once into the written work for each such interval. Accordingly, in the written work, the first and second differences have the same appearance as in the case of the ordinary parabola, but they are not exact theoretical values.

As previously mentioned, the second differences are then regarded as a function which is to be computed from the modified third and fourth differences in the same manner as a parabola is computed from first and second differences, except that the interval between values is three times as great. Each even-numbered order of modified differences is treated in the same way so that the interval between tabular entries is 9 for third and fourth orders, 27 for the fifth and sixth orders, and 81 for the seventh and eighth orders.

These deviations from the classical method cause a slight periodic deviation from the exact values of the original polynomial. The magnitude of this deviation is smaller the greater the number of points computed in a given interval. It can be predicted by formulas given graphically in Fig. 1 to determine whether it is larger than allowable in any particular case.

6. GETTING STARTED

The leading differences, that is, the top number in each column, must first be computed.

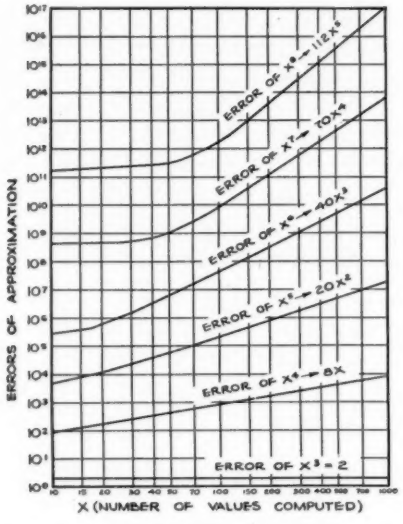

ERRORS OF APPROXIMATION
X (NUMBER OF VALUES COMPUTED)
ERROR OF X⁸ = 112X⁵
ERROR OF X⁷ = 70X⁴
ERROR OF X⁶ = 40X³
ERROR OF X⁵ = 20X²
ERROR OF X⁴ = 8X
ERROR OF X³ = 2

FIG. 1. The maximum error of approximation is found by algebraically adding the product of the coefficient of each term of the polynominal and the corresponding error shown here.

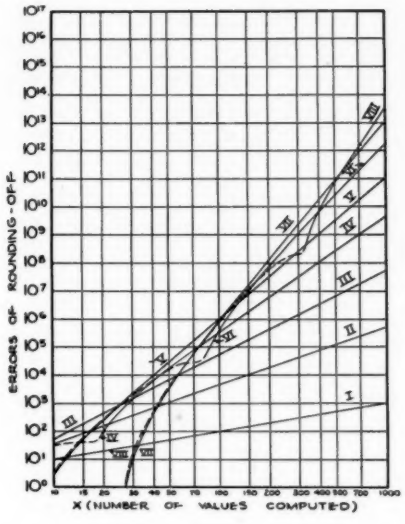


FIG. 2. The error of rounding-off is shown due to a +1.00 error of rounding-off in the designated leading difference. Broken lines indicate negative values.

Suppose the function

$$Y=A+BX+CX^2+DX^3+EX^4 +FX^5+GX^6+HX^7+JX^8$$

is to be computed for $X=$successive integers, beginning with zero. The formulas in Table III were derived by a long and laborious process (see Appendix) and by them the computer determines each leading difference, using only as many significant figures as is judged necessary.

These formulas are here given with all significant figures so that in extremely close work the rounding-off errors may be reduced to any limit desired.

7. NUMERICAL EXAMPLE

As a numerical example, the equation

$$\begin{aligned} Y=&\underset{(A)}{5.6875}+\underset{(B)}{9.6875X\cdot 10^{-2}}+\underset{(C)}{0.625X^2\cdot 10^{-4}} \\ &\underset{(D)}{-5.625X^3\cdot 10^{-6}}-\underset{(E)}{2.8125X^4\cdot 10^{-8}} \\ &\underset{(F)}{-0.5625X^5\cdot 10^{-10}} \end{aligned}$$

has been computed from $X=0$ to $X=100$. The leading tabular values as found from the formulas in Table III are as follows:

	As computed	As rounded off
$Z_{V,13}$	$-0.5467500\cdot 10^{-6}$	$-0.54675\cdot 10^{-6}$
$Z_{IV,0}$	$-0.0607500\cdot 10^{-4}$	$-0.06075\cdot 10^{-4}$
$Z_{III,4}$	$-1.104485625\cdot 10^{-4}$	$-1.10448\cdot 10^{-4}$
$Z_{II,0}$	$+0.012516875\cdot 10^{-2}$	$+0.012516\cdot 10^{-2}$
$Z_{I,1}$	$+9.704594274\cdot 10^{-2}$	$+9.70460\cdot 10^{-2}$
Y_0	$+5.68750$	$+5.68750$

The reader is asked to pass over for the moment the peculiar kind of rounding-off of decimals, the purpose of which is to illustrate errors of rounding-off, and which will be discussed in more detail later.

The beginning and the ending of the tabular computation are shown in Table IV. This much was considered sufficient to illustrate the method.

8. ERRORS OF APPROXIMATION

As mentioned before, this method for higher degree curves is not rigorously exact, but the approximation is amply close for most work. The magnitude of the error can be predicted by adding algebraically the errors pertaining to the separate terms of the original polynomial. These errors are shown graphically in Fig. 1 assuming a constant coefficient of unity for each order. The value found from the graph is, therefore, to be

multiplied by the corresponding coefficient of the polynomial.

A threshold value is shown in the graph below which lower order error terms predominate.

Every value of BX and CX^2 is exact, as is every third value of DX^3 and EX^4, every ninth value of FX^5 and GX^6, and every twenty-seventh value of HX^7 and JX^8. Above the threshold value the second computed point after a theoretically exact value has a larger error than the other intervening points in the case of EX^4, FX^5, and GX^6, while this is true of the twentieth point in the case of HX^7 and JX^8. Again above the threshold, if $(X-1)$ is a multiple of 3, the computed value is larger than the theoretical value, while if $(X-2)$ is a multiple of 3, it is smaller.

In the numerical example already given, we find the following errors corresponding to $x=100$:

Maximum error in DX^3 is less than	-0.000011
Maximum error in EX^4 is less than	-0.000023
Maximum error in FX^5 is less than	-0.000012
Maximum error of approximation	-0.000046

9. ERRORS OF ROUNDING-OFF

In addition to the theoretical errors inherent in the system, errors are introduced by the rounding-off of decimals. For practical purposes, these may be assumed to be simply the rounding-off errors in the leading differences. It is true that subsequent values are rounded-off also, but for two reasons the effect is smaller. In the first place, the error has less time to accumulate in repeated additions, and in the second place, the plus and minus errors tend to cancel each other. By substituting the final value of X into the equations in Table V, also shown graphically in Fig. 2, one finds the error in the final value of Y due to a $+1.00$ error in the given leading difference. These are then multiplied by the actual rounding-off error in each case and added together algebraically to obtain an estimate of the total final rounding-off error. Or conversely, by assuming a maximum allowable error in the final Y, the allowable rounding-off errors in the leading differences can be determined.

Applying these formulas to our numerical example and assuming the maximum allowable error in the final Y to be 0.00005 in each case, we arrive at the following:

Leading difference	(Final error)/(initial error)	Allowable initial error
Z_{I}	100	$\pm 0.00000\ 05$
Z_{II}	4850	$\pm 0.00000\ 0010$
Z_{III}	55540	$\pm 0.00000\ 00009$
Z_{IV}	379750	$\pm 0.00000\ 00001\ 3$
Z_{V}	1020200	$\pm 0.00000\ 00000\ 5$

In the numerical example, the leading differences were purposely rounded-off the wrong way so as to illustrate the effect. The actual prediction of the final error is as follows:

Leading difference	Actual initial error	Predicted final error
Z_{I}	$+0.00000\ 005726$	$+0.000006$
Z_{II}	$-0.00000\ 000875$	-0.000042
Z_{III}	$+0.00000\ 000056$	$+0.000031$
Error of rounding-off:		-0.000005
Error of approximation, previously given:		± 0.000046
Total predicted error:		$+0.000041$ or -0.000051
Actual final error:		$+0.00003$ and -0.00006

These figures agree to a useful degree with the actual results of the numerical computation as shown in Table IV. Exact agreement could not be expected on account of neglecting all the rounding-off errors after the initial ones.

10. CHANGING COORDINATES TO FIT THE FORMULA

Since X has been defined as taking on the values of the successive integers, it is necessary to transform most equations. The alternative of generalizing the formulas has been decided against as leading to greater complications. Although the necessary procedure is well known, it is illustrated here for ready reference.

TABLE V. Errors due to rounding-off of decimals.

Leading difference	Approximate error in final Y due to $+1.0$ rounding-off error
$Z_{\text{I, }1\frac{1}{2}}$	X
$Z_{\text{II, }0}$	$X^2/2-3X/2$
$Z_{\text{III, }4\frac{1}{2}}$	$X^3/18-X/6$
$Z_{\text{IV, }0}$	$X^4/216-X^3/12+X^2/72+X/4$
$Z_{\text{V, }13\frac{1}{2}}$	$X^5/9720-17X^3/1944+13X/90$
$Z_{\text{VI, }0}$	$X^6/524880-X^5/6480+5X^4/34992+17X^3/1296-29X^2/1620-13X/60$
$Z_{\text{VII, }40\frac{1}{2}}$	$X^7/99202320-19X^5/1180980+4051X^3/524880-3571X/3780$
$Z_{\text{VIII, }0}$	$X^8/2\ 14277\ 01120-X^7/661\ 34880+91X^6/15305\ 50080+57X^5/23\ 61960-18979X^4/1133\ 74080-4051X^3/3\ 49920+59639X^2/8\ 16480+3571X/2520$

Instead of the numerical example in the form given previously, suppose that points were to be computed on the curve

$$U=0.25V^2-0.00018V^5,$$

every 0.05 from $V=5$ to $V=10$. Horner's method of transposing coordinates so that the origin is at $V=5$ goes by successive synthetic divisions:

−0.00018	0.0000	0.0000	0.25	0.00	0.0	5
		−0.0009	−0.0045	−0.0225	+1.1375	5.6875
−0.00018	−0.0009	−0.0045	+0.2275	+1.1375	+5.6875	
		−0.0009	−0.0090	−0.0675	+0.80	
−0.00018	−0.0018	−0.0135	+0.1600	+1.9375		
		−0.0009	−0.0135	−0.135		
−0.00018	−0.0027	−0.027	+0.025			
		−0.0009	−0.018			
−0.00018	−0.0036	−0.045				
		−0.0009				
−0.00018	−0.0045					

so that the transposed equation

$$U=5.6875+1.9375W+0.025W^2 -0.045W^3-0.0045W^4-0.00018W^5,$$

where $W=(V-5)$ is now to be computed from $W=0.0$ to $W=5$ mm.

Now, since W increases in steps of 1/20 mm, then $W=(X/20)$ for X to increase in steps of 1. Then

$$U=Y=5.6875+1.9375\frac{X}{20}+0.025\frac{X^2}{20^2} -0.045\frac{X^3}{20^3}-0.0045\frac{X^4}{20^4}-0.00018\frac{X^5}{20^5}$$

$$=5.6875+9.6875X\cdot10^{-2}+0.625X^2\cdot10^{-4} -5.625X^3\cdot10^{-6}-2.8125X^4\cdot10^{-8} -0.5625X^5\cdot10^{-10}.$$

This is the form for computing, and the procedure from this point on is as previously shown.

11. CURVE OF nTH DEGREE THROUGH $(n+1)$ POINTS

In many cases the easiest approach is to determine a number of values of the function and to compute a smooth curve through these points. In the general case, the computer has the choice of setting up simultaneous linear equations in A, B, C $\cdots$, and solving, or of using the less familiar Lagrange formulas. A streamlined method for solving simultaneous linear equations has been described by Crout.[4] As soon as the coefficients of the equation in X are known, the computer proceeds by using Table III as described earlier. It is understood, of course, that the scale of X is so chosen that, in the table to be computed, X increases in steps of unity.

In the special case in which the $(n+1)$ known points are equally spaced in X, if one of them can be chosen as the origin, a more direct method is to compute the differences to the nth order in the classical manner. The symbol Δ^* is now introduced as denoting either a difference lying on the horizontal line through the value chosen as the origin or the arithmetical mean of a pair of difference values lying a half-space above and below this line. It may be necessary to extend the difference array upward or downward by adding or subtracting differences in order to find the Δ^* values. Now, taking M as a constant such that the known points are for $X=-mM$, $\cdots$, $-M$, 0, M, $2M$, $\cdots$, and $(n-m)M$, the Z values can be computed by the equations in Table VI.

It should be pointed out that M need not be an integer. Incidentally, while the mathematicians may have a word for it, the ratio represented by M is, to an optical designer, merely a magnification of the scale interval of X.

12. ADAPTING TO HIGHER ORDERS OR CLOSER WORK

For computing equations of still higher degree, or in some few cases in which the errors of approximation would be too large, it is sometimes practical to derive an equation[5] for the first difference $(Y_{(X+1)}-Y_X)$, where

$$Y_X=A+BX+CX^2+\cdots$$

and

$$Y_{(X+1)}=A+B(X+1)+C(X+1)^2+\cdots,$$

or perhaps for the second or a higher order difference. This derived equation can then be computed by the short-cut method exactly as heretofore described, after which the summation of differences proceeds by the classical method until the original polynomial is computed.

---

[4] P. D. Crout, Trans. A. I. E. E. 1235–41 (1941).

[5] Quickly done by using a table of binomial coefficients. See Scarborough, reference 2, pp. 45–46.

TABLE VI. Leading differences for computing nth degree curve through $(n+1)$ equi-spaced points.

$$Z_{\text{VIII},0} = 5\ 31441 M^{-8}\Delta^*_{\text{VIII}}$$

$$Z_{\text{VII},40\frac{1}{2}} = 19683 M^{-7}\Delta^*_{\text{VII}} + \frac{3}{2}Z_{\text{VIII},0}$$

$$Z_{\text{VI},0} = \frac{729}{M^6}\left[\Delta^*_{\text{VI}} - \left(\frac{1}{4} - \frac{91}{4M^2}\right)\Delta^*_{\text{VIII}}\right]$$

$$Z_{\text{V},13\frac{1}{2}} = \frac{81}{M^5}\left[\Delta^*_{\text{V}} - \left(\frac{1}{3} - \frac{38}{M^2}\right)\Delta^*_{\text{VII}}\right] + \frac{3}{2}Z_{\text{VI},0}$$

$$Z_{\text{IV},0} = \frac{9}{M^4}\left[\Delta^*_{\text{IV}} - \left(\frac{1}{6} + \frac{5}{2M^2}\right)\Delta^*_{\text{VI}} + \left(\frac{7}{240} + \frac{5}{8M^2} + \frac{1\ 75361}{80M^4}\right)\Delta^*_{\text{VIII}}\right]$$

$$Z_{\text{III},4\frac{1}{2}} = \frac{3}{M^3}\left[\Delta^*_{\text{III}} - \left(\frac{1}{4} - \frac{17}{4M^2}\right)\Delta^*_{\text{V}} + \left(\frac{7}{120} - \frac{17}{12M^2} - \frac{29999}{40\ M^4}\right)\Delta^*_{\text{VII}}\right] + \frac{3}{2}Z_{\text{IV},0}$$

$$Z_{\text{II},0} = \frac{1}{M^2}\left[\Delta^*_{\text{II}} - \left(\frac{1}{12} + \frac{1}{4M^2}\right)\Delta^*_{\text{IV}} + \left(\frac{1}{90} + \frac{1}{24M^2} + \frac{1069}{40M^4}\right)\Delta^*_{\text{VI}} - \left(\frac{1}{560} + \frac{7}{960M^2} + \frac{1069}{160M^4} + \frac{1764\ 64907}{2240M^6}\right)\Delta^*_{\text{VIII}}\right]$$

$$Z_{\text{I},1\frac{1}{2}} = \frac{1}{M}\left[\Delta^*_{\text{I}} - \left(\frac{1}{6} - \frac{1}{2M^2}\right)\Delta^*_{\text{III}} + \left(\frac{1}{30} - \frac{1}{8M^2} - \frac{383}{40M^4}\right)\Delta^*_{\text{V}} - \left(\frac{1}{140} - \frac{7}{240M^2} + \frac{383}{120M^4} - \frac{99\ 54067}{560M^6}\right)\Delta^*_{\text{VII}}\right] + \frac{3}{2}Z_{\text{II},0}$$

Table VI is intended for use in fitting a single continuous nth-degree curve to $(n+1)$ points. A closer approximation can often be obtained by using this particular curve only between two neighboring anchor points and using a different nth-degree curve in the next interval. This procedure is common in the standard methods of interpolation. Further work is in progress by which it is hoped to adapt the short-cut technique to the latter procedure also.

APPENDIX

The computing method which has been described was developed gradually and with much groping around. At first, it was only thought of as applying to an ordinary parabola, or to the two final orders of differences of a higher degree function. Means were then sought to extend the full advantages of the method to the cubic and to higher degree equations. The cubic was not too bad, as an exact computation could be made by writing one-third of the first differences and two-thirds of the second differences. (These were $\Delta_{\text{I},1\frac{1}{2}}$, $\Delta_{\text{II},1}$, and $\Delta_{\text{II},2}$.) The third difference, of course, was constant and did not need to be written except at the top of the page. Higher orders, however, led to a routine that was too complicated to be practical.

Then one day while doing an example of a cubic or a quartic, it seemed a waste of time to write the two nearly equal values of Δ_{II} between each two succeeding values of Δ_{I}. Would any appreciable error be introduced by using the average of the two in place of both? A trial showed that there would be a cumulative error, but it was soon apparent that this error could be counteracted by adjusting the values of the leading differences. Trials for the best approximation showed (as could be expected) that the error could be completely removed for only one value out of three if any continuing and regular routine of computing was adhered to.

After making the approximation just mentioned, it was a comparatively easy step to see that the same routin could be applied to computing the modified second dife ferences from some as yet not exactly known third and fourth differences, the corresponding entries in the table being spaced three times as far apart. Similarly, the scheme could be extended indefinitely to the fourth and sixth orders and so on. There remained, however, a long and tedious process of defining exactly the different quantities involved and of ascertaining whether the errors introduced would be prohibitive.

Derivation of Tables V and III

After reaching its final form, the computing method was extended to (at present writing) eighth-degree polynomials and the earlier derivations were repeated by computing a set of functions which will be referred to as primitive functions. These primitive functions were computed by setting the leading modified difference of each order in turn equal to unity while all the other leading terms were set equal to zero. The third- and fourth-order primitive functions are defined by the computed values of Y for $X=0, 3, 6, 9, \cdots$, similarly the fifth- and sixth-order for $X=0, 9, 18, 27, \cdots$, and the seventh- and eighth-order for $X=0, 27, 54, 81, \cdots$, since these sets of values can be freed from errors of approximation. The eight primitive functions are shown in full in Table V, where they do double duty as also showing the rounding-off errors. It was found easier to compute these primitive functions directly than to try to define them from a theoretical analysis of the necessary conditions.

Knowing that the commutative and distributive laws apply, it follows that multiples of these primitive functions can be combined to form the simple functions $Y=X^n$. For

example:

$$\begin{aligned} 216\times 4\text{th order} &= X^4-18X^3+3X^2+54X \\ +324\times 3\text{rd order} &= \quad +18X^3 \quad -54X \\ -6\times 2\text{nd order} &= \quad -3X^2+9X \\ -9\times 1\text{st order} &= \quad -9X \\ \text{Total} &= X^4. \end{aligned}$$

Thus, multiplying through by the coefficient E, it is seen that if $Z_{IV,0}=216E$; $Z_{III,4}=324E$; $Z_{II,0}=-6E$; and $Z_{I,4}=-9E$; then the term EX^4 would be (approximately) computed.

In a similar manner the conditions were determined for each of the eight terms of the polynomial and combined into the formulas shown in Table III.

Derivation of Table VI

The derivation of the formulas in Table VI was simple in plan, but tedious to carry out in detail. The coefficients of an eighth-degree polynomial through nine anchor points were determined in terms of the classical array of differences using central differences according to Stirling's interpolation formula. The coefficients of the polynomial were then eliminated by using the formulas of Table III.

JOURNAL OF THE OPTICAL SOCIETY OF AMERICA VOLUME 33, NUMBER 9 SEPTEMBER, 1943

Simplified Formulae for Scattered and Rescattered Sunlight*

LUDWIK SILBERSTEIN
Eastman Kodak Company, Rochester, New York

(Received May 3, 1943)

THIS subject, under the assumption of a flat earth or plane-stratified atmosphere, was treated four years ago by A. Hammad and S. Chapman.[1] Their formulae for the reception of the primary scattered sunlight are simple enough, but those for the reception of rescattered or "secondary scattered" light are, in their own words, "immensely more complicated," so much so that they are scarcely suitable for a numerical evaluation or even a general estimate of the effect. To see this, a glance on p. 109[1] will suffice.

The purpose of the present paper is merely to simplify these formulae by certain obvious reductions and by discarding unimportant terms, and thus to make them easily applicable. The merit of the method will be entirely due to these two authors. We shall retain, for the present, all their assumptions and, on the whole, their notation also.

Let $\rho=\rho(h)$ be the density of air at a level h above the ground and $M=\int_0^\infty \rho dh$, the total mass of air over unit area. Then it is convenient to introduce, instead of h, the variable

$$m=\frac{1}{M}\int_0^h \rho dh \qquad (1)$$

or its supplement, $\bar{m}=1-m$. The former is the fraction of the mass of air below, and the latter the fraction above the level h.

Now, consider any two points, P at a level h and P' at a level h'. Let r be their distance apart and θ the angle between their join and the vertical. Then $h'-h=r\cos\theta$ or $r=(h'-h)\sec\theta$. Now, with a view to the sequel, it is convenient to deal with $\int_P^{P'}\rho dr$ as a kind of effective distance between these points. By (1), this may be written

$$\int \rho dr = M\sec\theta\int dm = M\sec\theta\cdot(m'-m).$$

The absorption and scattering by any volume element of air are assumed to be proportional to *the mass* of that element. For a given wavelength, let α and σ be the absorption and the scattering coefficients of air per unit mass, and let us put

$$a=M\alpha, \quad s=M\sigma, \quad \text{and} \quad c=a+s.$$

Then the attenuation exponent of light flux in passing from P to P', due to the loss by absorption and scattering, will be

$$(\alpha+\sigma)\int \rho dr = c\sec\theta(m'-m)=\kappa(m'-m),$$

where $\kappa=c\sec\theta$, so that, with the inclusion of spreading by the inverse square law, the flux

---

* Communication No. 929 from the Kodak Research Laboratories.

[1] A. Hammad and S. Chapman, Phil. Mag. **28**, 99–110 (1939), where references to earlier investigations will be found.

received per unit area at the point $P'(m')$ will be

$$I'=\frac{I}{r^2}\exp\left[-\kappa(m'-m)\right], \qquad (2)$$

where I is the flux per unit solid angle radiated from the point $P(m)$ and r is the distance from P to P'. The variable $\kappa(m'-m)$ is thus, in general, a most convenient substitute for geometrical distance, which is constantly used by Hammad and Chapman. It fails, however, when the two points are on the same level. For, in that case, $m'-m=0$ and $\kappa=\infty$. We may, therefore, avail ourselves of the formula (2), keeping in mind, however, that for every horizontal transit ($h=$const.) the indeterminate product $\kappa(m'-m)$ is to be replaced by its original meaning,

$$(\alpha+\sigma)\rho r=\frac{c\rho}{M}r,$$

where ρ is the air density at the contemplated level. This clause will scarcely entail any inconvenience.

Refraction due to density variations being neglected, all light paths are here taken to be rectilinear. Polarization, also, is wholly disregarded. All symbols of intensities or fluxes, as well as of the optical coefficients, are intended to refer to some narrow spectral region and to denote their amounts per unit wave-length interval.

Let z be the zenith distance (angle) of the sun and I_0 the flux of sunlight per unit area "at the top of the atmosphere," that is, before appreciably penetrating into it, and let

$$Z=c\sec z.$$

Then the flux of direct sunlight per unit area at a level h or $m=1-\bar{m}$ is, by (2), since the variation of the spreading factor is negligible,

$$I=I_0\exp\left[-Z\bar{m}\right], \qquad (3)$$

e.g., at the ground, $I=I_0e^{-Z}$. This formula would give, at sunrise, or sunset, $I=0$, which is palpably wrong, the discrepancy being due to the assumption of a flat earth. In fine, its application cannot be pushed just up to $z=90$ degrees, though it may be accurate enough for zenith angles but little smaller than a right angle.

By the definition of σ, the light scattered by the particle thus illuminated, per unit mass, in all directions is σI, and thus, in a direction $\mathbf{k}$ making with the incident beam an angle ψ, per unit solid angle, by Rayleigh's law,

$$E_1(h,\mathbf{k})=A\sigma I(1+\cos^2\psi),$$

where A is a numerical constant. Its value is readily determined. If $d\tilde{\omega}$ is an infinitesimal solid angle, we have $\int E_1 d\tilde{\omega}=\sigma I$, and since $\int(1+\cos^2\psi)d\tilde{\omega}=4\pi+(4/3)\pi$, the formula becomes, by (3),

$$E_1(h,\mathbf{k})=\frac{3\sigma I_0}{16\pi}(1+\cos^2\psi)\exp\left[-Z\bar{m}\right]. \qquad (4)$$

If θ, ϕ are the polar coordinates of the direction (unit vector) $\mathbf{k}$, the former counted from the zenith and the latter from the vertical plane through the zenith and the sun, then

$$\cos\psi=\cos z\cos\theta+\sin z\sin\theta\cos\phi.$$

E.g., with the sun in zenith, $\cos\psi=\cos\theta$.

With this meaning of the symbols, formula (4) gives the *emission of primary scattered* sunlight in any required direction $\mathbf{k}(\theta,\phi)$ from a particle placed at the level $h(\bar{m})$, per unit mass. Since $\cos^2(\pi-\psi)=\cos^2\psi$, the emission in the opposite direction $(-\mathbf{k})$ has the same value.

Now, for the reception of the sunlight thus scattered; it will suffice to develop its expression for the case in which the receiving element of area dS is placed at the ground level. Let $\mathbf{k}$ be its normal and $R_1(\mathbf{k})d\tilde{\omega}dS$ the primary scattered light striking it in directions contained within the elementary solid angle $d\tilde{\omega}$. All this light will be due to the air contained within the cone $d\tilde{\omega}$ with $\mathbf{k}$ as axis and apex P placed at dS. The mass of an element of this cone, at a distance r from P, being $\rho r^2 dr d\tilde{\omega}$, its emission (flux per unit solid angle) is $E_1\rho r^2 dr d\tilde{\omega}$. By (2), the flux per unit area received at dS is $E_1\rho dr d\tilde{\omega}e^{-\kappa m}$, so that

$$R_1(\mathbf{k})=\rho\int E_1 e^{-\kappa m}dr=M\sec\theta\int_0^1 E_1 e^{-\kappa m}dm$$

or, by (4), and since $\bar{m}=1-m$,

$$R_1(\mathbf{k})=\frac{3sI_0}{16\pi c}\kappa(1+\cos^2\psi)e^{-Z}\int_0^1 e^{(Z-\kappa)m}dm.$$

Ultimately, therefore, *the primary scattered radiation received* per unit area at the ground level within the cone $d\tilde{\omega}$ in any direction or, as we may say, from any "point" (θ, ϕ) of the sky, is $R_1(\theta, \phi)d\tilde{\omega}$ in which

$$R_1(\theta, \phi)=\frac{3sI_0}{16\pi c}\kappa(1+\cos^2\psi)\frac{e^{-\kappa}-e^{-Z}}{Z-\kappa}, \qquad (I)$$

where $\cos\psi=\cos z\cos\theta+\sin z\sin\theta\cos\phi$ and $Z=c\sec z$, $\kappa=c\sec\theta$.

This is Hammad and Chapman's formula, when the receiving point or element is placed at $h=0$, namely, their formula for Rd_1 (*down*-going light), p. 104, in this special case.

For $\theta=z$, and any ϕ,

$$R_1(z, \phi)=\frac{3sI_0}{16\pi c}Ze^{-z}(1+\cos^2\psi),$$

where $\cos\psi=1-\sin^2 z(1-\cos\phi)$. Thus the primary scattered radiation received *per unit solid angle* in the direction straight from the sun $(\phi=0)$,

$$R_1(z, 0)=\frac{3sI_0}{8\pi c}Ze^{-z},$$

while the direct sunlight received is, as above,

$$R_0(z, 0)=I_0e^{-z}.$$

For comparison of the brightness of the sky with that of the sun, the scattered radiation may be computed for a cone subtending the same solid angle as the sun ($\epsilon=0.000066$):

$$\frac{\epsilon R_1(z, 0)}{R_0(z, 0)}=0.000066\left(\frac{3s}{8\pi}\right)\sec z.$$

Since the scattering coefficient s is (for $\lambda=\frac{1}{2}\mu$) of the order of $\frac{1}{10}$, this ratio is of the order of $8\times10^{-7}\sec z$, e.g., 1.6×10^{-6} for $z=60$ degrees and about 10^{-5} for $z=85$ degrees. But, for reasons just given, the formula can scarcely claim complete reliability at such high angles, that is, for such low positions of the sun.

If the sun is *in zenith*, $z=0$, $Z=c$, $\psi=\theta$, and the expression (I), now of course symmetrical around the sun, reduces to

$$\frac{R_1(\theta)}{R_1(0)}=\frac{1}{2c}\cdot\frac{1+\cos^2\theta}{1-\cos\theta}[1-e^{-c(\sec\theta-1)}], \qquad (I_0)$$

where

$$R_1(0)=\frac{3sI_0}{8\pi}e^{-c}.$$

For small angles, even up to $\theta=20$ or 25 degrees,

$$\frac{R_1(\theta)}{R_1(0)}=\frac{1+\cos^2\theta}{2\cos\theta}.$$

A numerical evaluation of these simple formulae will be given later on.

Now, for the much more intricate determination of the rescattered or *secondary* scattered sunlight.

Since our main purpose, in the present connection, is to find out the relative importance of the secondary and the primary scattering, it will suffice to consider the case of *the sun in zenith*, i.e., $z=0$, $Z=c$.

An air particle P placed at the level h (or m) receives primary scattered light from all directions, namely, the amount $R_1(h, \mathbf{k}')d\tilde{\omega}'$ from every direction $\mathbf{k}'(\theta', \phi')$ within the solid angle $d\tilde{\omega}'=\sin\theta' d\theta' d\phi'$, of which it rescatters, per unit mass, in any fixed direction $\mathbf{k}(\theta, 0)$,* the fraction $(3\sigma/16\pi)(1+\cos^2\chi)$, where

$$\cos\chi=\mathbf{kk}'=\cos\theta\cos\theta'+\sin\theta\sin\theta'\cos\phi',$$

$\mathbf{kk}'$ being the scalar product of the two unit vectors. Thus, the complete emission from $P(h)$ of secondary scattered light in the direction $\mathbf{k}$, per unit mass and unit solid angle, is

$$E_2(h, \mathbf{k})=\frac{3\sigma}{16\pi}\int(1+\cos^2\chi)R_1(h, \mathbf{k}')d\tilde{\omega}'.$$

The expression for R_1 is to be constructed as before, only with $h(m)$ replacing the ground level and m', ψ', κ' written instead of m, ψ, κ, so that

$$R_1=\frac{3sI_0}{16\pi}(1+\cos^2\psi')\sec\theta' \times\int_0^1\exp\left[-Z\bar{m}-\kappa'|m'-m|\right]dm',$$

where $|m'-m|$ stands for the absolute value of

* Since, with the sun in zenith, everything is symmetrical about the vertical, it is enough to consider $\mathbf{k}$ in a single vertical plane and we may count the longitude ϕ from just this plane.

$m'-m$. Since, in our case, $Z=c$, $\psi'=\theta'$, and $\bar{m}'=1-m'$,

$$R_1=\frac{3sI_0}{16\pi}(1+\cos^2\theta')\sec\theta'\cdot e^{-c}\times\int_0^1\exp\left[cm'-\kappa'|m'-m|\right]dm'.$$

The required emission of rescattered sunlight thus becomes

$$E_2(h,\mathbf{k})=\frac{9s\sigma I_0}{256\pi^2}e^{-c}\int_0^{\pi/2}\int_0^{2\pi}\int_0^1\times\exp\left[cm'-\kappa'|m'-m|\right]\tan\theta'(1+\cos^2\theta')\times(1+\cos^2\chi)d\theta'd\phi'dm'.$$

The angle ϕ' occurs only in $\cos^2\chi$, as defined above, and we have

$$\frac{1}{\pi}\int_0^{2\pi}(1+\cos^2\chi)d\phi'=2+2\cos^2\theta\cos^2\theta'+\sin^2\theta\sin^2\theta'.$$

Next, the integral over m' is

$$\int_0^m\exp\left[cm'-\kappa'(m-m')\right]dm'+\int_m^1\exp\left[cm'-\kappa'(m'-m)\right]dm'=e^c\left\{\frac{\exp\left[-c\bar{m}\right]-\exp\left[-\kappa'\bar{m}\right]}{\kappa'-c}+\frac{\exp\left[-c\bar{m}\right]-\exp\left[-(\kappa'm+c)\right]}{\kappa'+c}\right\}\equiv e^cF(\theta'),\text{ say.}$$

Thus,

$$E_2(h,\mathbf{k})=\frac{9s\sigma I_0}{256\pi}Q, \tag{5}$$

where

$$Q=\int_0^{\pi/2}F(\theta')\{2+2\cos^2\theta\cos^2\theta'+\sin^2\theta\sin^2\theta'\}(1+\cos^2\theta')\tan\theta'd\theta'$$

or, with $x=\sec\theta'$ as integration variable,

$$Q=\int_1^\infty F(x)\left\{2+\sin^2\theta+(4-2\sin^2\theta)\frac{1}{x^2}+(2-3\sin^2\theta)\frac{1}{x^4}\right\}\frac{dx}{x}.$$

If we introduce the coefficients

$$A_0=2+\sin^2\theta,\quad A_1=4-2\sin^2\theta,\quad A_2=2-3\sin^2\theta, \tag{6}$$

depending only on θ, then

$$Q=A_0\int_1^\infty F(x)\frac{dx}{x}+A_1\int_1^\infty F(x)\frac{dx}{x^3}+A_2\int_1^\infty F(x)\frac{dx}{x^5}. \tag{7}$$

It remains to evaluate the three integrals in terms of the level (m) of the emitting particle, say

$$L_n(m)=\int_1^\infty F(x)\frac{dx}{x^n},\quad n=1,3,5, \tag{8}$$

where

$$cF(x)=\frac{\exp\left[-c\bar{m}\right]}{x-1}+\frac{\exp\left[-c\bar{m}\right]}{x+1}-\frac{\exp\left[-c\bar{m}x\right]}{x-1}-\frac{\exp\left[-c(mx+1)\right]}{x+1}. \tag{9}$$

The integral $L_1(m)$ is easily evaluated. The first two terms of $F(x)$ lead to logarithms and the last two to the exponential integral as defined by

$$Ei(-x)=\int_\infty^x\frac{e^{-u}du}{u}.$$

Collecting all terms, we find without difficulty

$$cL_1(m)=\exp\left[-c\bar{m}\right]\left[Ei(\epsilon)-\log\epsilon+\log 2c\bar{m}-Ei(-2c\bar{m})\right]-Ei(-c\bar{m})+e^{-c}Ei(-cm),$$

where $\epsilon\to 0$. Now, $Ei(\epsilon)-\log\epsilon$ is Euler's constant, $C=0.5772\cdots$, so that, ultimately,

$$cL_1(m)=\exp\left[-c\bar{m}\right]\left[C+\log 2c\bar{m}-Ei(-2c\bar{m})\right]-Ei(-c\bar{m})+e^{-c}Ei(-cm). \tag{10}*$$

The remaining two integrals, $L_3(m)$ and $L_5(m)$, may be derived from $L_1(m)$ by means of a recurrence formula given by Hammad and Chapman.

In the first place, however, let us simplify the rigorous expression (10) for $L_1(m)$. Since c is of the order of $\frac{1}{10}$ and m, $\bar{m}\leq 1$, we may retain of the well-known expansion of Ei the first three terms only,

$$Ei(x)=C+\log|x|+x,$$

* This is identical with the last formula of §17 of Hammad and Chapman's paper for $L_1(m)$, with $z=0$, $Z=c$.

and reject also the higher terms in the expansion of the exponentials. Thus we find

$$L_1(m)=0.423-m\log cm-\bar{m}\log c\bar{m}, \quad (10a)$$

where the constant term is $1-C$, to three decimals.

Hammad and Chapman's recurrence relation,* given at the end of §17,[1] amounts in our case ($z=0$, $Z=c$) to

$$cL_3(m)=cL_1(m)+Ej_3(c\bar{m})+Ej_2(c\bar{m})$$
$$-e^{-c}Ej_3(cm)+e^{-c}Ej_2(cm)-2\exp[-c\bar{m}]$$

and

$$cL_5(m)=cL_3(m)+Ej_5(c\bar{m})+Ej_4(c\bar{m})$$
$$-e^{-c}Ej_5(cm)+e^{-c}Ej_4(cm)-\tfrac{2}{3}\exp[-c\bar{m}],$$

where

$$Ej_1(x)=-Ei(-x)$$

and, for $n>1$,

$$Ej_n(x)=\frac{1}{n-1}[e^{-x}-xEj_{n-1}(x)],$$

so that all required Ej_n are at once reduced to Ei. Substituting the values of the numerous terms thus defined and rejecting again all c^2 terms, we are left with $cL_3(m)=\frac{1}{2}c$, $cL_5(m)=\frac{1}{4}c$, so that the two integrals are constants, up to c terms,

$$L_3(m)=\tfrac{1}{2}, \quad L_5(m)=\tfrac{1}{4}. \quad (11a)$$

So also is $L_1(m)$, (10a), correct up to c terms only.

With the functions $L_n(m)$ just defined, the emission of secondary light, (5) with (7) and (8), becomes

$$E_2(h,\mathbf{k})=\frac{9s\sigma I_0}{256\pi}\{A_0L_1(m)+A_1L_3(m)+A_2L_5(m)\}. \quad (12)$$

It contains the direction $\mathbf{k}(\theta)$ of the emission through the A's, as given by (6), and the level h or m of the emitting particle through the $L_n(m)$, in the present approximation through $L_1(m)$ only. The reception at the ground level of the light thus rescattered will be expressed, in just the same way as the primary reception, by

$$R_2(\theta)=M\sec\theta\int_0^1 E_2e^{-\kappa m}dm.$$

* This relation should follow from our definition, (8) with (9), of the three $L_n(m)$.

Thus, and since $M\sigma=s$, the formula for *the reception of the secondary scattered* sunlight becomes

$$R_2(\theta)=\frac{9s^2I_0}{256\pi}\sec\theta\{A_0L_1+A_1L_3+A_2L_5\}, \quad (II)$$

where

$$L_n=\int_0^1 e^{-\kappa m}L_n(m)dm.$$

By (11a), the integrals L_3 and L_5 can be written down directly. It remains to evaluate the integral L_1 with $L_1(m)$ as given by (10a). This can be accomplished rigorously in terms of elementary functions and the function Ei, with the aid of a simple integration by parts.

We have, by (10a),

$$L_1=\frac{0.423}{\kappa}(1-e^{-\kappa})-S-T,$$

where

$$S=\int_0^1 e^{-\kappa m}m\log cm\cdot dm,$$

$$T=\int_0^1 e^{-\kappa m}\bar{m}\log c\bar{m}\cdot dm.$$

Now,

$$\kappa^2S=\int_0^\kappa xe^{-x}\log\left(\frac{cx}{\kappa}\right)dx$$

$$=-\left|\log\frac{cx}{\kappa}\cdot(x+1)e^{-x}\right|_\epsilon^\kappa+\int_\epsilon^\kappa\left(1+\frac{1}{x}\right)e^{-x}dx$$

$$=1-e^{-\kappa}+\int_\epsilon^\kappa\frac{e^{-x}dx}{x}+\log\epsilon$$
$$+\log\frac{c}{\kappa}-\log c\cdot(1+\kappa)e^{-\kappa},$$

where $\epsilon\to 0$. Thus, and recalling that $Ei(-\epsilon)-\log\epsilon=C$,

$$\kappa^2S=0.423+\log\frac{c}{\kappa}$$
$$-\{1+(1+\kappa)\log c\}e^{-\kappa}+Ei(-\kappa).$$

Next, since $m+\bar{m}=1$,

$$T=-e^{-\kappa}\int_1^0 e^{\kappa\bar{m}}\bar{m}\log c\bar{m}\cdot d\bar{m}=\frac{e^{-\kappa}}{\kappa^2}\int_0^\kappa xe^x\log\frac{cx}{\kappa}\cdot dx,$$

whence, integrating by parts and reasoning as before,

$$\kappa^2 T=\left\{0.423+\log\frac{c}{\kappa}+Ei(\kappa)\right\}e^{-\kappa}+(\kappa-1)\log c-1.$$

This settles L_1, while L_3 and L_5 follow directly from (11a).*

Collecting the results, we have the following approximate values of the three functions L_n:

$$L_1=\frac{0.423}{\kappa}(1-e^{-\kappa})-S-T,$$

$$L_3=\frac{1}{2\kappa}(1-e^{-\kappa}), \qquad (13)$$

$$L_5=\frac{1}{4\kappa}(1-e^{-\kappa}),$$

where

$$\kappa^2 S=0.423+\log\cos\theta+Ei(-\kappa)-\{1+(1+\kappa)\log c\}e^{-\kappa}$$
$$\kappa^2 T=\{0.423+\log\cos\theta+Ei(\kappa)\}e^{-\kappa}+(\kappa-1)\log c-1 \qquad (14)$$

and $\kappa=c\sec\theta$. This holds for any angle θ.

For small angles, κ is but slightly greater than c and, therefore, of the order of $\frac{1}{10}$. E.g., the mean of the Mount Wilson values, for $\lambda=\frac{1}{2}\mu$, is $c=0.15$. Thus, for small θ, up to some 8 or 10 degrees,

$$S=T=\tfrac{1}{2}\left\{\log c-\frac{\sin^2\theta}{c^2}-\tfrac{1}{2}\right\}$$

and

$$L_1=0.923+\frac{\sin^2\theta}{c^2}-\log c,\quad L_3=\tfrac{1}{2},\quad L_5=\tfrac{1}{4}. \qquad (13_0)$$

E.g., for $c=0.15$ and $\theta=0$, $L_1=2.82$, a moderate number. For $\sin\theta\doteqdot c$ or θ equal to about $8\frac{1}{2}$ degrees, $L_1=3.82$ is still a moderate number. At the other extreme, for $\theta=\pi/2$, the formulae give zero for all three L_n, as would follow also directly from their definition,

$$L_n=\int_0^1 e^{-\kappa m}L_n(m)\,dm.$$

With the functions L_n determined by (13), (14) and the A's expressed by (6), the reception of the rescattered sunlight, with the sun in zenith, is now given by formula (II). For the sake of comparison with the reception of primary scattering it will be convenient to consider the ratio of $R_2(\theta)$ to $R_1(0)$. Since $R_1(0)=3sI_0/8\pi e^c$, the required expression becomes

$$\frac{R_2(\theta)}{R_1(0)}=\frac{3se^c}{32}\sec\theta\{A_0L_1+A_1L_3+A_2L_5\}. \qquad (II')$$

This may now be compared directly with the expression (I_0) for $R_1(\theta)$, the reception of primary scattering. (Sun in zenith for both.)

To form an idea of the relative importance of the contributions due to the primary and the secondary scattering in various directions (θ), let us assume the Mount Wilson attenuation coefficient mentioned earlier, $c=0.15$. The scattering coefficient $s=M\sigma$ is but slightly smaller than $c=s+a$. If, say, $s=0.12$, then the numerical factor in (II') becomes

$$\frac{3se^c}{32}=0.013,$$

whence we can judge that the secondary reception will be only a small or moderate fraction of the primary one.

Let us consider this comparison in some numerical detail.

The computation of the primary reception by means of the formula (I_0) is simple enough. The results, for $c=0.15$, are collected in the following table, in which we have included the $\pi/2$-value, $1/2c=10/3$, although the validity of the formula cannot be claimed for angles approaching 90 degrees.

$\theta=$	90°	86	85	80	75	70	60	50	45	40	30°
$\frac{R_1(\theta)}{R_1(0)}=$	3.333	3.113	2.916	2.120	1.676	1.419	1.161	1.055	1.031	1.014	1.001

* We may mention here that if Rayleigh's scattering factor $1+\cos^2\chi$ is dropped, for which, however, there are no good reasons, then $A_0=$ const. and $A_1=A_2=0$, so that we get rid of the functions L_3 and L_5. But $L_1(m)$ and its integral L_1 remain exactly as before, so that in the end not much is gained by such an omission.

Below 30°, $R_1(\theta) : R_1(0) = 1$, to three decimals.

On the other hand, we have, for the reception of the secondary scattered sunlight,

$$\frac{R_2(\theta)}{R_1(0)} = 0.013 \sec\theta\{A_0L_1 + A_1L_3 + A_2L_5\},$$

where

$$A_0 = 2+\sin^2\theta, \quad A_1 = 4-2\sin^2\theta, \quad A_2 = 2-3\sin^2\theta,$$

as in (6), and L_1, L_3, L_5 are given, to our approximation, by (13) and (14).

For $\theta = 0$, as we saw, $L_1 = 2.82$, $L_3 = \frac{1}{2}$, $L_5 = \frac{1}{4}$, so that $R_2(0)/R_1(0) = 0.106$ or a little above one-tenth.

For $\theta = 30$ degrees, when $\kappa = 0.173$,* we find, to three figures, $S = -1.53$, $T = -1.63$, whence $L_1 = 3.55$, $L_3 = 0.459$, $L_5 = 0.230$, and $R_2(\theta) : R_1(0) = 0.148$.

Again for $\theta = 60$ degrees the ratio turns out to be 0.190 and for $\theta = 80$ degrees, $R_2(\theta) : R_1(0) = 0.230$.

Comparing these numbers with those of the preceding table, we have the ratio of the two receptions:

at	$\theta =$ 0	30°	60°	80°
$R_2(\theta) : R_1(\theta) =$	0.106	0.148	0.164	0.109.

In fine, the contribution due to secondary scattering is from about one-tenth up to one-sixth of that due to primary scattering.

It seems likely that the ratio of the tertiary to the secondary contribution will be the same or even smaller. If so, then the emission and the reception of thrice-scattered sunlight, whose computation would be extremely tedious, may be entirely left out of account. It would, no doubt, be more important to take account of the neglected sphericity of the earth and the refraction of light in its atmosphere. Some minor, though perhaps not quite negligible, corrections may also be made by pushing our approximation for the functions $L_n(m)$ and L_n one step higher. (See *Note* at the end of the paper.) In the meantime, it would be interesting to test the formulae (I) and (II), as they stand, by actual observations on the brightness of the sky in various directions, with any position of the sun for the first, at least, and with a high position of the sun for both formulae.

* $Ei(\kappa) = -1.013$, $Ei(-\kappa) = -1.355$.

Note. Since $L_1(m)$ is, generally, some six and twelve times greater than $L_3(m)$ and $L_5(m)$, it may suffice to consider the effect of the higher terms upon the first of these functions. In simplifying the rigorous expression (10) for $cL_1(m)$, we have retained only the terms of the order of c. To push the approximation one step higher, let us include all terms in c^2. Let $c\Delta L_1(m)$ be their total contribution to $cL_1(m)$. Then $\Delta L_1(m)$ itself will be of the order of c, namely, after simple reductions,

$$\Delta L_1(m) = c\Big\{\tfrac{1}{3}C - 1 + \tfrac{1}{4}(m^2+3\bar{m}^2) + m + \tfrac{1}{2}m^3\log\frac{\bar{m}}{m} + \tfrac{1}{2}\log cm\Big\}.$$

The corresponding correction of L_1, as required in the final formula (II) or (II'), is $\Delta L_1 = \int_0^1 e^{-\kappa m}\Delta L_1(m)\,dm$ and, as before, might be evaluated without difficulty. This, however, would greatly encumber our working formula. To have a substantial estimate of the correction, it may suffice to consider the $m = \frac{1}{2}$-value of $\Delta L_1(m)$ which is

$$\Delta L_1(\tfrac{1}{2}) = \frac{c}{2}\Big(0.08 + \log\frac{c}{2}\Big),$$

while

$$L_1(\tfrac{1}{2}) = 0.423 - \log\frac{c}{2}.$$

Thus, e.g., with $c = 0.15$,

$$\Delta L_1(\tfrac{1}{2}) = -0.19, \quad \text{while} \quad L_1(\tfrac{1}{2}) = 3.01.$$

In round figures, $\Delta L_1(\frac{1}{2}) = -1/15 \cdot L_1(\frac{1}{2})$. Such, more or less, would also be the reduction of L_1. The values of L_3 and L_5 are, as we saw, of lesser importance. Ultimately, therefore, the inclusion of these terms would reduce the reception $R_2(\theta)$ by about 7 percent only of its values given in the text of the paper, and the effect of yet higher terms would, to all purposes, be evanescent.